EVOLUTION

Simulating Human Origins and Evolution

Ken Wessen

CSBEA

CAMBRIDGE

CAMBRIDGE

more information - www.cambridge.org/9780521843997

Cambridge Studies in Biological and Evolutionary Anthropology 42

Simulating Human Origins and Evolution

The development of populations over time and, on longer time scales, the evolution of species are both influenced by a complex of interacting, underlying processes. Computer simulation provides a means of experimenting within an idealised framework to allow aspects of these processes and their interactions to be isolated, controlled and understood.

In this book, computer simulation is used to model migration, extinction, fossilisation, interbreeding, selection and non-hereditary effects in the context of human populations and the observed distribution of fossil and current hominoid species. The simulations described enable the visualisation and study of lineages, genetic diversity in populations, character diversity across species and the accuracy of reconstructions, allowing new insights into human evolution and the origins of humankind for graduate students and researchers in the fields of physical anthropology, human evolution and human genetics.

KEN WESSEN has Ph.D.s in both Theoretical Physics and Human Evolution and has worked as a post-doctoral researcher in Computer Visualisation. He currently works in quantitative finance, and is an Adjunct Lecturer in the School of Anatomy and Human Biology at the University of Western Australia.

Cambridge Studies in Biological and Evolutionary Anthropology

Series editors

Also available in the series

21 *Bioarchaeology* Clark S. Larsen 0 521 65834 9 (paperback)
22 *Comparative Primate Socioecology* P. C. Lee (ed.) 0 521 59336 0
23 *Patterns of Human Growth*, second edition Barry Bogin 0 521 56438 7 (paperback)
24 *Migration and Colonisation in Human Microevolution* Alan Fix 0 521 59206 2
25 *Human Growth in the Past* Robert D. Hoppa & Charles M. FitzGerald (eds.) 0 521 63153 X
26 *Human Paleobiology* Robert B. Eckhardt 0 521 45160 4
27 *Mountain Gorillas* Martha M. Robbins, Pascale Sicotte & Kelly J. Stewart (eds.) 0 521 76004 7
28 *Evolution and Genetics of Latin American Populations* Francisco M. Salzano & Maria C. Bortolini 0 521 65275 8
29 *Primates Face to Face* Agustín Fuentes & Linda D. Wolfe (eds.) 0 521 79109 X
30 *Human Biology of Pastoral Populations* William R. Leonard & Michael H. Crawford (eds.) 0 521 78016 0
31 *Paleodemography* Robert D. Hoppa & James W. Vaupel (eds.) 0 521 80063 3
32 *Primate Dentition* Daris R. Swindler 0 521 65289 8
33 *The Primate Fossil Record* Walter C. Hartwig (ed.) 0 521 66315 6
34 *Gorilla Biology* Andrea B. Taylor & Michele L. Goldsmith (eds.) 0 521 79281 9
35 *Human Biologists in the Archives* D. Ann Herring & Alan C. Swedlund (eds.) 0 521 80104 4
36 *Human Senescence – Evolutionary and Biocultural Perspectives* Douglas E. Crews 0 521 57173 1
37 *Patterns of Growth and Development in the Genus* Homo. Jennifer L. Thompson, Gail E. Krovitz & Andrew J. Nelson (eds.) 0 521 82272 6

38 *Neanderthals and Modern Humans – An Ecological and Evolutionary Perspective* Clive Finlayson 0 521 82087 1

39 *Methods in Human Growth Research* Roland C. Hauspie, Noel Cameron & Luciano Molinari (eds.) 0 521 82050 2

40 *Shaping Primate Evolution* Fred Anapol, Rebecca L. German & Nina G. Jablonski (eds.) 0 521 81107 4

41 *Macaque Societies – A Model for the Study of Social Organization* Bernard Thierry, Mewa Singh & Werner Kaumanns (eds.) 0 521 81847 8

Simulating Human Origins and Evolution

K. P. WESSEN
University of Western Australia

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo

Cambridge University Press
The Edinburgh Building, Cambridge CB2 2RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521843997

First published in print format 2005

ISBN-13 978-0-511-11153-2 eBook (NetLibrary)
ISBN-10 0-511-11153-3 eBook (NetLibrary)

ISBN-13 978-0-521-84399-7 hardback
ISBN-10 0-521-84399-5 hardback

Each change of many-coloured life he drew
Exhausted worlds, and then imagined new.

Samuel Johnson 1709–84

For Cindy, Jessamine and Xanthe

Contents

Preface

Recent times have seen a great deal of activity and progress in human origins research, from the advent of molecular methods in the 1960s to the many important fossil hominid discoveries of the past few years. Nevertheless, the debate over whether particular fossil species are direct human descendants or not, and whether the fossil record and molecular results support a recent African origin or multiregional continuity, continues to rage. There is clearly a substantial need for fundamental work studying the methods employed in the interpretation of these data. The primary aim of the research presented in this volume is to begin to address this need by means of direct computer modelling and simulation of the many underlying and interacting processes.

Specifically, this volume describes the development and application of two related, but distinct, simulations, each designed to model important aspects of evolution in general, and the origin and evolution of humans in particular, as well as to provide substantial analysis and a wide variety of visualisations of the results.

The first simulation, *Specialist*, models the evolution of species and subspecies over millions of years, by starting with a single ancestral species with a particular suite of morphological 'characters' and allowing it to evolve in discrete steps. The characters are either hereditary or non-hereditary, and at each step a small number of these characters may change, either owing to random mutation or as a result of a change in the nature of the home environment of each species. Random extinction, fossilisation, interbreeding of subspecies, migration between four continents, and selective advantage are included in the model.

The main focus is on using the resulting species data to construct a phylogeny and migration history, which is then compared against the known true phylogeny. Two techniques of reconstruction are employed. The first technique involves matching existing species and fossils to the most closely (morphologically) related earlier fossil, whereas the second involves a reconstruction based on differences between the characters of the existing species only.

The second simulation, *Genie*, models several generations of individuals in up to three independent populations, thus allowing study of the effects of different mating patterns, fertility, adult sex ratio, migrations of various types, limiting population size, selective advantage and the impact of external,

natural disasters on common ancestry and the mixing of lineages generally. Once a complete genealogy is generated, common ancestry, lineage mixing and migrations are determined and analysed for the purely paternal and maternal genealogies (corresponding to Y chromosome and mitochondrial DNA inheritance), as well as the *biological* genealogy, or pedigree, where lineages are traced back through both parents simultaneously. This analysis is carried out on both a small sample of individuals and the full population, and individuals in the population carry both sex-specific and autosomal genes that are subject to mutation and recombination in controlled ways.

The simulations presented are essentially simulations of evolutionary change, and as such may be applied across a very large range of problems. As is apparent from the title of this volume, I have chosen to focus on problems relating to human and hominoid evolution, but extensions to many other areas are relatively straightforward, especially for the species/subspecies simulation.

Both simulations can do either single runs, with various visualisations and interactions, or multiple runs, with more limited visualisation but with basic statistical analysis of the results and all the required information for more advanced analysis provided in a simple text report. In particular, the simulated demographic, genetic and genotype data from *Genie* may be easily exported into other programs to provide more detailed or custom analysis. This removes the need for *Genie* to try to cover the myriad of possible analyses.

The programs that implement the simulations may be freely downloaded by following the links from http://school.anhb.uwa.edu.au/personalpages/kwessen.

In the interests of quality control, some minor limitations have been placed on the downloadable versions of the software, but these limitations can be removed via a simple registration process that will also allow me to provide updates and maintain some degree of dialogue with users. It is my hope that making the software available will lead to much further and diverse development of the simulations in collaboration with other researchers.

In addition, many of the figures in this book are black and white, or otherwise adjusted, versions of colour visualisations produced by the simulations. For this reason, the majority of simulations presented in the text are available for download along with the software, enabling them to be viewed in colour, and also enabling the many interactions provided by the simulation program to be explored in the context of these particular simulations. In order to gain a full appreciation of the results presented in this book, readers are urged to download the associated software and familiarise themselves in a hands-on way with the models and visualisations employed.

I take this opportunity to heartily thank Professor Charles Oxnard, without whose ongoing encouragement and highly infectious enthusiasm this project would never have begun, let alone finished. I also thank Professor Paul O'Higgins for his comments on an earlier version of this work; those comments were particularly instrumental towards providing the necessary impetus for me to undertake the publication of this work in book form. Professor Colin Groves also had several useful comments on an earlier manuscript, and various suggestions from Algis Kuliukas have led to valuable enhancements to the species simulation. Thanks are also due to Mat Abdy for his help in preparing this book's associated website.

And, of course, my most sincere and personal thanks go to my beautiful wife Cindy and lovely daughters Jessamine and Xanthe, each of whom will, I'm sure, very much share my relief at seeing this book complete!

1 *Introduction*

When considering the natural world, it is impossible not to be astounded at the extraordinary diversity of species it contains, and such feelings can only be magnified by the further realisation that what we are seeing is merely a 'snapshot' of the four thousand million year history of life on this planet. Understanding the generation of such a complex situation seems almost totally beyond comprehension, and, indeed, in many respects it is; but, like wonder and astonishment, curiosity is also a fundamental human trait, and through the efforts of many remarkable individuals, significant insight into the source and development of this diversity has been achieved.

Most notable was the discovery by the young Charles Darwin, travelling as a biologist on the *Beagle* in the early 1830s, that all species are related by common descent, and that the vast diversity observed is simply a product of the accumulation of small, but favourable, modifications over enormous periods of time. However, there remained the problem of explaining the inheritance of these modifications, since any form of inheritance where features in offspring are some kind of average of parental features would simply lead to dilution and eventual loss of favourable mutations. A solution was eventually provided by the work of the Austrian monk Gregor Mendel. In the 1860s, Mendel studied the inheritance of various external features of the pea plant in an environment of controlled cross-pollination, and was able to show that the inheritance of characters proceeded in a discrete fashion, with some characters dominant and some recessive. This work, rediscovered in 1900, laid the foundations for the field of genetics, and when fully integrated with evolutionary theory by Fisher, Haldane and Wright in 1930, in what is known as the *modern synthesis*, the major gap in Darwin's theory was filled. Since this time, the theory of evolution has provided a solid framework for understanding the generation of the diversity of species, and continues to grow in strength as the primary unifying thread in modern biology.

1.1 Phylogenetics and human origins

Darwin's reluctance to draw the conclusions from his theory relevant to human origins publicly is well known; his only mention of human ancestry in

The Origin of Species (1859) is a single sentence. But even this brief mention was sufficient to spark substantial conflict with those who saw his theory as an affront to orthodox religious belief. Fortunately, Darwin had an able assistant in his 'bulldog', Thomas Huxley, whose work, *Evidence as to Man's Place in Nature* (1863), confronted the issue of human origins head on.

The Swedish naturalist Linnaeus had already classified humans in the order Primates in the eighteenth century, but it was the work of Darwin, Huxley and others that led to the recognition that the similarities between humans and the so-called great apes – gorillas, chimpanzees, and orangutans – indicated common ancestry; most probably fairly recent common ancestry. Within an evolutionary framework of common descent, the problem of determining relatedness becomes one of constructing a *phylogeny*, i.e. a hypothesised evolutionary history, showing ancestry and branching at those points where mutation has led to a new species.

The *morphological* approach to determining a phylogeny is based on the study of particular physical characteristics of individuals across related species, e.g. size, shape and number of teeth, various body structures and bone lengths, and attempting to deduce the pattern of *speciation* events required to describe the observed differences. A particular strength of this approach is that fossil data, when available, can be included directly. In general, human–ape (i.e. hominoid) phylogenies so constructed had the great apes as one group (family Pongidae), and humans as a sister taxon (family Hominidae) (Oxnard, 1997; Simpson, 1945).[1]

It wasn't until the 1930s that an understanding of the biological mechanism of Mendelian inheritance began to emerge, culminating in 1951 when James Watson and Francis Crick demonstrated the double-helix structure of the DNA molecule, explaining at the same time the manner of its replication, and the way in which errors in this replication naturally lead to mutations. The amazing nature of DNA is apparent in many respects: it is self-replicating, it uses a genetic code that is essentially identical across *all* species, and it carries not only all the coding necessary for the development of a particular individual, but also a record of the evolutionary history of the species of that individual. Theoretical and practical advances in molecular genetics in recent years have allowed unprecedented access to this genetic information, and more and more is able to be read by using a variety of direct and indirect methods.

Early applications of these new approaches to the study of hominoid evolution were by Goodman in the early 1960s (Goodman, 1995), and, in particular,

[1] In the light of later developments, it is interesting to note that Huxley (1863) and Darwin (1871) both considered humans most closely related to the African apes (the gorilla and chimpanzee); see Mann and Weiss (1996) for a historical overview.

by Sarich and Wilson in 1967 (Sarich and Wilson, 1967). Sarich and Wilson employed an immunological technique, measuring the cross-reaction of antigens and antibodies from different hominoid species, as a method of comparing amino acid sequences, the degree of cross-reaction being a measure of similarity. The immune system is obviously highly important in natural selection, and therefore the results obtained by using this method are strongly correlated with the evolution of the species being studied. Results from this new research revealed the fact that humans, chimpanzees and gorillas are in fact more closely related to each other than any of them is to orangutans, so a more accurate phylogeny groups humans, gorillas and chimpanzees (the African apes) together, with orangutans as a sister taxon (see Figure 1.1). The ground-breaking aspect of this work was the imposition of a time scale, leading to an estimate of the time of the human–chimpanzee common ancestor of around 5 million years ago, far more recent than was being indicated by other work at the time.

Other molecular methods include direct sequencing of DNA, comparison of DNA sequences by using DNA–DNA hybridisation (Ruvolo, 1996, 1997), sequencing of various proteins, e.g. fibrinopeptides, haemoglobins and myoglobins (Jones *et al*., 1991), and comparison of the number, shape and banding patterns of chromosomes, i.e. the *karyotype* (Jones *et al*., 1991). These have all revealed a similar picture, although there are inconsistencies with respect to the order of the human–gorilla–chimpanzee split, and placing the orangutan with the African apes, or with the other Asian apes (i.e. the gibbon and siamang). Whole-organism morphological studies, and some recent soft-tissue morphological studies, have since been shown to agree with the molecular consensus (Collard and Wood, 2000; Oxnard, 1997).

All molecular approaches rely on direct access to an organism's DNA or living cells, and therefore, except in rare cases, cannot be applied to fossil species.

By classifying particular morphological features as *characters*, or by working directly with discrete genetic data, a formal approach to the study of evolutionary relationships is possible by using *cladistics* (Hennig, 1966), a method of study that provides a rigorous framework for the construction of phylogenies. Given a set of species, a set of characters are defined by which these species are to be compared, and an *outgroup* species identified, i.e. a closely related species, outside of the group and equally related to all members of the group. Character states that are present in the common ancestor of the group are classified as *primitive*, and do not provide any means of discrimination within the group. Identification of these states is based on the nature of the outgroup species, available fossil evidence, and their simplicity and commonality. Character states that are a result of change within the group are

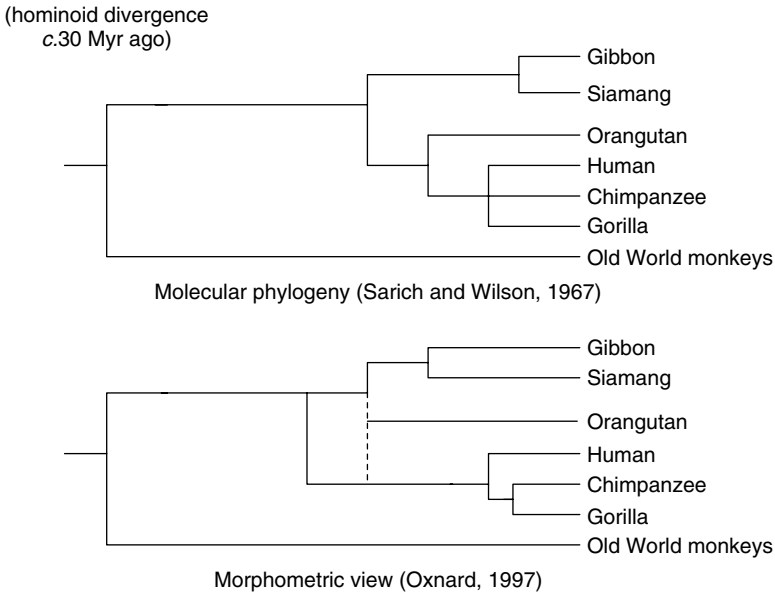


Figure 1.1. Sarich and Wilson's hominoid phylogeny (Sarich and Wilson, 1967), based on an immunological approach, is shown as a representative molecular phylogeny. Below this is the result of Oxnard's whole-organism morphological study (Oxnard, 1997). The most notable differences are that the morphological data are inconclusive as to whether or not the Asian apes form a clade (as indicated by the dashed line in the figure), and the relative timing of the human–chimpanzee–gorilla split. An approximate linear time scale spanning 30 million years can be applied to the molecular phylogeny. The morphological phylogeny is drawn so as to indicate a rough correspondence in this respect, although morphological data do not allow imposition of a time scale in the same way as do molecular data.

classified as *derived*, and may be shared within the group as a result of either homology, i.e. common ancestry, or homoplasy, i.e. similarity due to either independent mutation (*convergence* or *parallelism*) or reversion to a primitive state (*reversal*). Homoplasy acts to obscure phylogenetic relationships by giving a false appearance of shared ancestry, and so it is the identification of shared derived characters due to homology that leads to the construction of the most likely phylogeny. A group of species comprising a common ancestor and all its descendants is known as a *clade*, and is the fundamental unit of a cladistic classification.

Whatever approach is employed, morphological or molecular, there are a number of intrinsic limitations that must be recognised and, if possible,

overcome (Cronquist, 1987; Sokal, 1985). There are fundamental problems with the raw data. Morphological methods are most limited by the fact that there is no simple and reversible mapping between observable features and the underlying genetic programming, and by the fact that some characters reflect non-hereditary development, e.g. via the effects of biomechanics on bones, joints and muscles, i.e. external effects due to individual activity patterns (although these may be selected when of genetic origin). In more detail, hereditary character states may be the result of random mutations, or selected hereditary adaptations to some long-continued factor such as climate change, or be functionally adaptive (i.e. a selected hereditary character). Non-hereditary characters may also be the result of some randomness (e.g. due to disease), or be directed (e.g. in ontogenetic response to continued undernutrition), or be functionally adaptive, i.e. produced during ontogeny by interaction with the home environment of each species.

Furthermore, treating any particular character (and its associated states) as representative of some single evolutionary entity that may be meaningfully classified as primitive or derived for phylogenetic purposes, as described above, is a drastic oversimplification. Each morphological character, i.e. observable feature, is a complex of potentially quite a large number of underlying characters, and these are, of course, a mixture of various types. It is therefore not even theoretically possible to unambiguously classify the character under consideration. Further complications arise because of interdependencies between the underlying characters, meaning that, even if desired, it is difficult to quantify the contributions of the underlying characters in a way that preserves the ability to compare across characters. Oxnard (2000) discusses specific techniques for identifying parallels and convergences in morphological data, and thus avoiding confusion in phylogenetic interpretations.

For example, the character ‘cranial length’, considered primitive in apes when long and derived in humans when short, provides a useful illustration of this problem. Cranial length actually comprises several lengths, each resulting from a mix of hereditary or non-hereditary effects with varying degrees of uncertainty. Firstly, the length of the outer table of the skull at the front is probably related to load-bearing and may depend upon both heredity (thicker skulls may run in a family) and function during life (non-hereditary) (such as powerful chewing of skins). Next is the thickness of the diploe (i.e. the marrow cavity within the skull bone at the front), which may be related to blood-forming (probably hereditary), and perhaps also to biomechanics with the bone operating in a poroelastic rather than an elastic mode (probably non-hereditary). The length from front to back of the frontal air sinus is perhaps related to the physiology of the respiratory system and perhaps

also to stress-bearing; these characteristics could be either hereditary or non-hereditary. Then there is the length of the anterior fossa of the cranial cavity, which contains the frontal lobes of the brain but is also dependent upon the size of the eyeball, and next the length of the middle cranial fossa, containing the parietal lobes but also dependent on middle- and inner-ear adaptations. The length of the clivus is probably related to brain and skull growth, but also partly of somitic origin and therefore dependent on somite origin and growth. The length of the foramen magnum depends on the size of spinal cord, and thus on the input–output relationship between body and brain. The length of the occipital planum depends on the size of the posterior lobes of the brain, but maybe also on the strength of the nuchal muscles. Then there are three more measures relating to the inner table of the occipital bone, the diploe at the back and the outer table of the occipital bone.[2]

Obviously all these 'characters' could have, and should have, different cladistic designations. What, then, is to be made of the overall character of cranial length? Certainly a simplistic classification as either primitive or shared derived cannot capture anything like the complete picture.

Classification difficulties aside, the lack or incompleteness of relevant fossils, and problems in the identification of fossils, creates difficulties that must also be overcome. For example, no gorilla or chimpanzee fossils have been identified; all candidate fossils from the past 6 million years have been placed on the human line (Gee, 2001). Molecular methods are limited by the facts that DNA mutates at an unknown rate, is differentially selected based on its consequent selective advantage or disadvantage, is mixed each generation, and its transfer is restricted in various ways, e.g. owing to geographic constraints, migration and breeding patterns (Avise, 2000).

Once the data have been obtained and identified, applying the method described above also involves overcoming a number of practical complications. Identification of appropriate characters is the first problem. It is important that the characters used are independent of each other, and properly reflect evolutionary history and not individual adaptation. Once the characters have been identified, classification of their states as primitive or derived, and, if shared derived, homologous or homoplasious, is difficult and prone to error. Then there are a number of different methods for constructing the phylogeny, often producing no single best choice. The computational methods required are intrinsically hard,[3] meaning that there exists no reasonable (polynomial) time algorithm able to guarantee finding the best solution: as the number of

[2] See, for example, Oxnard (2004) for a related discussion.

[3] In mathematical terms, they belong to the family of *NP-complete* problems (Day *et al*., 1986; Graham and Foulds, 1982).

characters is increased, the search time required to locate the best phylogeny increases exponentially. Furthermore, results based on different measures are often inconsistent, and are sensitive to the ordering of characters, the order of input, the approximations employed to overcome limitations of the chosen algorithm, etc. (Felsenstein, 1982; Sokal, 1985).

Despite these problems, the general agreement regarding the evolutionary relationships between humans and apes that has emerged since the advent of molecular studies has proven quite robust.

1.2 Origin of modern humans

When fossils of archaic humans are included in the above picture, things become much less clear. Despite the general agreement on the evolutionary relationship of humans and apes, when it comes to details of the human lineage there is substantial disagreement over issues such as the rate of evolution, the number of distinct evolutionary lineages involved, the extent of interbreeding, and what migrations have occurred (Relethford, 1998), leading to serious contention in the matter of modern human origins.

The argument is usually presented as a conflict between two models (Smith and Harrold, 1997). According to the *Recent African Origin* or *Replacement* model, anatomically modern humans emerged as a new species in Africa around 200 000 years ago, and then spread throughout the Old World, replacing existing populations without significant interbreeding. Opposed to this is the *Multiregional Evolution* or *Regional Continuity* model, which views all human evolution as taking place within a single evolutionary lineage. According to this model, modern humans arose simultaneously everywhere, as a result of interregional gene flow. These two models are extreme positions on a spectrum of such models, each hypothesising various degrees of replacement and continuity.

A great deal of research involved in the attempt to distinguish between these possibilities concerns mitochondrial DNA. Mitochondria are small cellular organelles important for metabolism, and each contains a small (*c*. 16 000 bp) circular genome known as mitochondrial DNA (mtDNA). Mitochondrial DNA has two very important properties that make it extremely useful in the study of modern human origins. Firstly, it is purely maternally inherited (see Avise (2000) and Strauss (1999) for a discussion of paternal leakage); secondly, it mutates an order of magnitude faster than nuclear DNA (Seielstad *et al*., 1998), and therefore can resolve much shorter time scales. Polymorphisms in the mtDNA data allow the construction of a phylogeny, and given an estimate

of the effective population size, i.e. the number of breeding individuals over the period of interest, and an estimate of the mutation rate, a time may be assigned to the depth of the tree, and to the migrations therein.

Cann *et al*. (1987) studied mtDNA collected from 147 individuals of Asian, Australasian, European and African ancestry. The phylogeny they constructed had two branches, one of which consisted entirely of individuals with African ancestry, leading to the conclusion that the common mtDNA ancestor of all humans, the so-called *Mitochondrial Eve*, lived in Africa. This conclusion was further supported by the fact that the African populations showed the greatest variation in their mtDNA, an indication of greatest age. To determine the time of this common ancestor, an estimate of the mutation rate of mtDNA was required. Cann *et al*. obtained this by assuming a constant rate of mutation together with a date of 5 million years ago for the human–chimpanzee common ancestor. Given the present-day divergence of human and chimpanzee mtDNA, this led to an estimate of between 140 000 and 290 000 years ago, a date strongly in agreement with the Recent African Origin model. Many criticisms of the method employed have been addressed in later work (see, for example, Vigilante *et al*. (1991)).

The paternal analogue to mtDNA is provided by those parts of the Y chromosome that are not homologous to the X chromosome (Jobling and Tyler-Smith, 1995). Recent studies of Y-chromosome polymorphisms have mostly concurred with the mtDNA picture (Dorit *et al*., 1995; Hammer, 1995; Pritchard *et al*., 1999; Underhill *et al*., 1997; Whitfield *et al*., 1995) but have also acted to bring into focus the effect on the statistical analysis of the underlying demographic assumptions employed in these studies. Fu and Li (1996) reanalysed the results of Dorit *et al*. (1995) and showed that there is such a substantial dependence on the estimate of N, the effective population size, that it can change the estimate of the time of the most recent paternal common ancestor by an order of magnitude, with the mean ranging from 92 000 years for $N = 2500$ up to 703 000 years for $N = 30\,000$. Brookfield (1997) considers several such estimates, and concludes '... the estimates depend hardly at all on the data, and almost entirely on the demographic model assumed.'

There has also been the suggestion that mtDNA results imply a severe population bottleneck and that all modern humans are descended from an extremely small and recent founder population, even a single individual. Ayala (1995) addressed and thoroughly dismissed this suggestion, claiming that the data actually imply that the effective population never dropped below 100 000 individuals (although the size of this figure is now understood to be a result of balancing selection; see Sherry *et al*. (1998)). The consensus is for a population of the order of 10 000 breeding individuals (Gagneux *et al*., 1999; Relethford, 1998; Sherry *et al*., 1998), with evidence for a relatively

recent demographic bottleneck or selective sweep in human origins. It is interesting to note that for both mtDNA and the Y chromosome, the variation within humans, when compared with other primates, is surprisingly small, also implying a relatively recent divergence.

Similar work has been done using genes on autosomes, although the situation is far more complicated because of the two potential paths of inheritance each generation, and the corresponding greater depth of the resulting phylogenies. Harding *et al*. (1997) studied 326 sequences of the beta-globin gene, and found African common ancestry, dated at approximately 800 000 years ago, and no evidence of the effective population dropping below 10 000 individuals at any time. More significantly, they found a depth of greater than 200 000 years in their Asian sample, implying that the ancestral population was already widely dispersed at that time. A similar challenge to the Recent African Origin model comes from the analysis of the mtDNA of the *Mungo Man* (Adcock *et al*., 2001), an anatomically modern human found at Lake Mungo, Australia, and dated at about 60 000 years ago. It was found that despite being anatomically modern, his mtDNA lineage diverged before the most recent common ancestor of living human mtDNA. Relethford (1998) demonstrates how this entire class of results can be interpreted from a population perspective rather than from a phylogenetic perspective, and thus be shown to be consistent with both a recent African origin and multiregional evolution.

It must be remembered that a species tree is actually a combination of several individual gene trees, and the overall picture may only be recoverable through the study of several of these individual genes (Moore, 1995). The three species shown in Figure 1.2 contain a gene whose form in species C is older than the form in species A and B (the B–C species ancestor being polymorphic). Sampling this particular gene would incorrectly imply a closer relationship between species A and B than between B and C. (Analogously, in a morphological study, many independent morphological characters may be needed for accurate resolution of a species tree.)

There are problems not only with sampling effects and the underlying demographic assumptions as discussed above, but also with the assumptions regarding the other important input: the mutation rate.

As is apparent from the above discussion of the work by Cann *et al*. (1987), molecular methods rely on knowledge of the mutation rate of DNA across time and between species. The *molecular clock hypothesis* is a consequence of the neutral theory of evolution (Kimura, 1968) and implies an approximately constant rate of mutation, so long as the DNA sequence retains its original function. If this is the case, then the degree of difference between sequences being compared is simply proportional to the time since

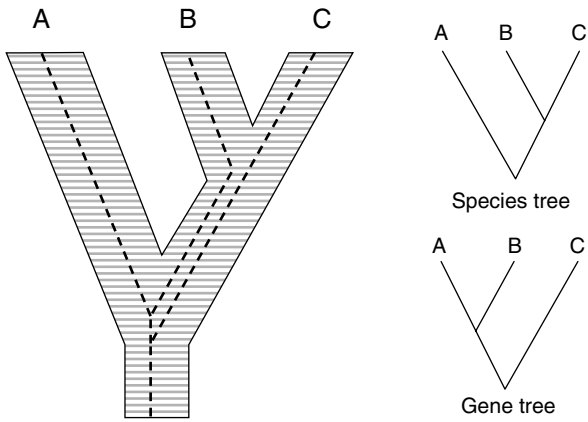


Figure 1.2. A phylogeny of three species, with the path of a particular gene shown by a dashed line. The simplified species tree and conflicting gene tree for these three sample species are shown on the right.

the sequences diverged. By incorporating fossil evidence, the *clock* can be calibrated, and thus divergence times can be attached to a molecular phylogeny.

In fact, particular DNA sequences and proteins can mutate at vastly different rates at different times and in different lineages, and although there may be some local validity of the molecular clock hypothesis, in general there is global failure (Avise, 2000; Gibbons, 1998; Ruvolo, 1996; Strauss, 1999; Wills, 1995). The fast-mutating microsatellite loci, i.e. short repetitive sections of DNA that lie between genes, have been used to construct an alternative method for timing lineages that does not rely on external calibration of the rate of molecular evolution (Goldstein *et al*., 1995). However, because of mutational saturation, nuclear microsatellites are only useful for timing relatively recent events. In particular, the deepest split in the human phylogeny can be recovered with such a method, but saturation will occur in less time than the five million years or more back to the human–chimpanzee common ancestor (Jorde *et al*., 1998).

This situation also affects substantially the common ancestor calculations described above. For example, Wills (1995) includes a variable mutation rate across mtDNA sites and obtains a range of 436 000 to 800 000 years ago for the mitochondrial common ancestor, depending on the date used for the human–chimpanzee common ancestor.

In general, the molecular data seem to support the replacement hypothesis, but when all the aforementioned caveats are considered, it remains far from

conclusive. The dates vary widely, depending on the method and assumptions employed. Furthermore, a recent African origin has difficulty with the observed continuity of regional morphological traits, especially outside of Europe, whereas the multiregional hypothesis has difficulty with the amount of gene flow required for its support, as well as with a number of aspects of the molecular data. Perhaps the only thing that is truly clear is that population size, breeding patterns, local geographic events, migrations and reproductive barriers present a severe challenge when it comes to interpreting these results (Lahr and Foley, 1998). So long as positions at both extremes in this debate consider themselves equally well supported by the same data, be it fossil or molecular, substantial further study into the basis of all these methods is obviously of great importance.

1.3 Computer methods in phylogenetics

One approach to this problem is to try to understand more quantitatively the dependencies of the results on the many assumptions that one is forced to make because of the incompleteness of the fossil data, lack of knowledge of the demographics, and other problems. Computer simulations provide this capability by allowing experimentation within an idealised framework, through which various aspects of the problem may be unambiguously isolated, visualised and understood.

In the study of human origins, two very different processes are of concern: one involving the origin of species, and the other involving migrations and other demographic factors within a population of a given species. Computers have been used in both these areas, as a means of studying speciation, by providing tools for reconstructing phylogenies from various kinds of morphological or molecular data, and as a means of studying gene flow and genetic drift in populations by using statistical models.

Raup *et al*. (1973) studied the generation of species lineages by modelling speciation as an equilibrium process of random lineage branching. All lineages stem from a common ancestor, and may continue in time, become extinct, or produce a new lineage by branching, with a probability based on the difference between the existing diversity and a predetermined equilibrium value. An algorithm for the automatic identification of clades was included, allowing study of the taxonomy of the resulting phylogeny. The simulations produced quite a variety of clade shapes, which were then compared with actual clades for the Reptilia. An important fact demonstrated by this work is that differences in evolutionary pattern do not *necessarily* imply an inherent

difference in the associated taxonomic groups: simulated groups evolving under identical constraints can behave very differently. Sepkoski and Kendrick (1993) used a similar model to simulate phylogenies. Employing exponential, logistic and mass-extinction diversification profiles, the resulting phylogenies were degraded in various ways (to model the effects of fossilisation, for example) and the information content remaining was analysed with respect to the 'true' phylogeny. Both these models can be generalised to allow the study of higher taxa, e.g. genus, family, etc. Nee *et al*. (1994) also used a similar approach to study the reconstruction of phylogenies, looking particularly at the role of lineages that become extinct.

Another important advance due to the use of computers comes from the availability of excellent programs for calculating phylogenies by using a variety of methods, such as PHYLIP (Felsenstein, 1993) and PAUP* (Swofford, n.d.). Together with increasingly powerful desktop computers, these programs have made it possible for researchers to work with much larger data sets while also trying many different methods and data arrangements, thus producing more complete analyses. Nei (1991) provides a review of the efficiencies of the various phylogenetic tree-making methods by simulating the evolution of DNA sequences and then applying different phylogenetic methods to the results. By comparing the results of each method with the actual sequence generation, the relative efficiency of the various methods in recovering the correct phylogeny is measured.

A major area of current computer simulation research involves statistical modelling of population genetics, looking in particular at gene flow and genetic drift, under various models of migration, mutation and effective population. Such simulations bear direct relevance to the calculation of common ancestry and migration history from molecular data phylogenies, and it is to be hoped that they will lessen the confusion in interpreting results such as those described in the previous section.

The work by Ayala mentioned earlier (Ayala, 1995; Ayala and Escalante, 1995) employed simulations of the change in allele frequency over time based on random mating of individuals, with a reproduction probability higher for heterozygotes than for homozygotes by some specified amount. By changing the degree of selective advantage, the relationship between the population size and the number of alleles that survive is studied, and the results applied to the analysis of current allelic diversity and its implications for previous effective population size and time of the most common ancestor as discussed in the previous section. Wollenberg and Avise (1998) simulated gender-specific genetic pathways, including migration and various degrees of proximity-based limitations on mating. Having simulated a population history in this manner, they looked at the sampling properties and found that demographic

factors did indeed impact significantly on the statistical links between the true and estimated genetic history of the population.

The study of a population is generally retrospective in nature, starting with a sample from an existing population and then attempting to describe the observed features in terms of the population's prior evolution. Results are then generalised from the sample to the entire population. Coalescent theory (Kingman, 1982b; Hudson, 1990) provides a probabilistic framework perfectly suited to this approach, and has therefore become an extremely important tool in population genetics over the past 20 years. In brief, coalescent theory describes the merging of lineages from a sample of a population as one goes backwards in time, to the point where only a single lineage remains, i.e. the common ancestor. It is particularly well suited to molecular data, and although the usual formulation is based on the neutral model (Kimura, 1968) and a single, randomly mating population of constant size, extensions to cover recombination (Hudson, 1983), population growth (Kuhner *et al.*, 1998), population subdivision (Hudson, 1990; Donnelly and Tavaré, 1995) and selection (Neuhauser and Krone, 1997) are well developed and are the subject of much ongoing research. A good review may be found in Fu and Li (1999).

Various important parameters leave traces in the pattern of lineage-merging that can be identified and understood. In particular, it is possible to estimate the mutation rate (Donnelly and Tavaré, 1995; Fu, 1994, 1998), test for selective neutrality (Tajima, 1989; Fu and Li, 1993) and estimate the ancestral population size and the age of the most recent common ancestor (Tavaré *et al.*, 1997; Fu and Li, 1996, 1997; Griffiths, 1999; Tang *et al.*, 2002), as well as the recombination rate (Wakeley, 1997) and migration rate (Beerli and Felsenstein, 1999).

Other advantages provided by coalescent theory include robustness with respect to the specific mating model employed, since such details only affect the time scale of the process (Kingman, 1982a), as well as the fact that simulations based on coalescent theory are far more efficient than traditional population simulation approaches because only those lineages that contribute to the final sample population need be considered. Coalescent theory is discussed in far more detail in Chapter 8.

The aim of the simulations presented in this book is to model the actual processes of species and population evolution, rather than to employ statistical approximations. Advantages of such an approach include fewer assumptions and broad simplifications, and a more direct mapping between reality and the simulation. Disadvantages include the fact that any model is a vast oversimplification, and we are forced to work with smaller simulations, both in time and in size. This is especially limiting for the population simulations.

Nevertheless, such simulations can provide an understandable and quantitative picture of how various complications interfere with, and thus compromise, phylogenetic and genealogical reconstruction.

Two associated programs are available for download.[4] The species and subspecies simulations are implemented in the program *Specialist*, which enables the modelling of evolution, extinction, migration, non-hereditary characters, interbreeding, selective advantage and cladistic analysis. By using *Specialist*, it is possible to see directly the impact of the modelled complications on the ease and accuracy of phylogenetic reconstruction; in Part I of this book *Specialist* is used to simulate scenarios relevant to ancient human origins.

A second program, *Genie*, a genealogy simulator, is also available. *Genie* allows modelling of a population while taking into account the effects of overall population size, breeding patterns, sex ratios, various kinds of migration, selective advantage, reproductive success and bottlenecks. Two independent genetics models are included: the first models genotypes based on two alleles at a single locus, allowing for selection, dominance, overdominance and mutation, whereas the second allows arbitrary mutation of two mitochondrial genes, two Y-chromosome genes, and two nuclear genes. In the latter case, recombination may also be included. In Part II of this volume, *Genie* is used to study the problem of tracing back genealogies and thus give insight into the difficulties of determining modern human origins.

[4] See http://school.anhb.uwa.edu.au/personalpages/kwessen, as mentioned in the Preface.

Part I
Simulating species

2 *Overview*

The concept of a 'species', despite its fundamental nature, has proven surprisingly difficult to pin down, primarily because an ideal species concept seemingly must satisfy a number of conflicting criteria. Hull (1997) gives a review of several possibilities, evaluating them with respect to generality, applicability and theoretical significance, but finds that no single concept is clearly superior to the others. Particular difficulties for any species concept arise because of polymorphism, clinal variation (e.g. gradual change across a large area leading to the situation where interbreeding between local subspecies can occur, but is not biologically possible between subspecies more geographically separated), and hybrid zones (where gene flow is possible, but is limited by geographic constraints).

Although there are several proposed definitions, the following two examples are broadly representative of the two most widely used general approaches.

Biological species concept: a group of organisms is a species if it consists of actually or potentially interbreeding individuals, and is reproductively isolated from other such groups (Mayr, 1969).

Phylogenetic species concept: a group of organisms is a species if it is the least inclusive monophyletic group definable by at least one autapomorphy (i.e. a derived character state exclusive to a particular taxon) (Mishler and Donoghue, 1982). Closely related to this is the *diagnostic species concept* (Cracraft, 1983), where the classification is based on character states that are fixed and not necessarily autapomorphic.

In practical terms, these two definitions are not as different as they at first seem. Both of them attempt to define a species essentially as an evolutionarily independent unit; in genetic terms, the biological species concept implies that gene flow *can* occur, whereas the phylogenetic species concept implies that gene flow *has* occurred.

As a precursor to his taxonomy of living primates, Groves (2001) also reviews a range of species concepts, with particular attention to the operational–theoretical distinction, and himself uses a phylogenetic concept.

However, both concepts have limitations. The biological species concept is unable to classify asexual species, and neither can it be applied to fossil species; it tends to be overly *lumpy* and results in groups larger than perhaps are desired; and it can also be argued that, because the ability to interbreed is a primitive trait, the biological species concept may result in grouping of species that are not actually closest genetic relatives. The phylogenetic species concept is limited in ways that are in many respects the flip-side of the above problems. It tends to be overly splitting, resulting in groups smaller than desirable; organisms may be grouped on characteristics of unclear biological relevance; and different species (according to this definition) may interbreed, leading to interspecies gene flow.

A species is really only a concept of evolutionary significance if viewed over time, but species concepts in general struggle with the temporal dimension. The biological species concept is a time-slice of an evolving lineage, but lineages can only be recognised in retrospect and thus knowledge of the future is required for an accurate classification of current species. One particular extension of the biological species concept that explicitly takes this into account is the *evolutionary species concept*, which is based on the premise that 'evolutionary species' evolve separately from other such lineages (Hull, 1997; Simpson, 1961). Phylogenetic species concepts are synchronous, showing only sister-group relationships.

Evolution can be defined as 'change in allele frequency over time'; the most important underlying processes are gene flow, genetic drift, mutation and natural selection. Precisely when a new species can be said to have arisen depends somewhat on the definition of species employed, but the mechanisms of speciation can be classified, primarily geographically, as follows (Avise, 2000). *Allopatric* speciation arises as a result of geographic isolation. For example, when migration or some chance event separates a population, the allele frequency in the new populations will generally not be the same as in the parent population, and will, from the time of separation onwards, develop independently. Furthermore, such events produce smaller populations, and so the relative importance of genetic drift increases. *Sympatric* speciation occurs when subspecies occupy the same area, but are reproductively separate because of different home environments. A third possibility, *parapatric* speciation, occurs when gene flow between neighbouring populations is interrupted because of an abrupt local change that affects the fitness of each group in the other region.

In the species simulation, presented in the next chapter, speciation is modelled by a random change of characters, and the lack of interbreeding population modelling breaks any direct connection to the above mechanisms. However, the subspecies simulation captures each of these processes to some

extent because interbreeding is modelled, and the chance of interbreeding is determined by effects that parallel these processes. In particular, allopatric events are modelled directly by using migration, or more precisely, the imposition of reproductive isolation. Sympatric events are modelled indirectly by associating an environment with each subspecies and prohibiting interbreeding between subspecies living in different environments. Parapatric events are indirectly modelled via the combined effects of migration and selective advantage.

As discussed in the introduction, being able to associate a time scale with a phylogeny is crucial. For fossil methods this is directly so, and for molecular methods it is equally important for calibration purposes. In molecular studies of modern human ancestry, it is very important to have an accurate picture of the time and nature of the most recent human–chimpanzee common ancestor. This provides important calibration information required for determining the time and location of the common ancestor of the modern human population, as well as valuable evidence to be considered in the classification of potential human fossils from this period. To be able to obtain this information, an understanding of the species that have led to modern humans and chimpanzees, especially since divergence, is needed.

Current knowledge of the fossil situation for hominoids indicates the existence of many more species in the past than at present, but, as mentioned in the introduction, no fossils at all have been identified as chimpanzee or gorilla ancestors (Gee, 2001). The current fossil picture for hominoids and humans is described in detail in the next two sections.

Any group of species may be classified according to the phylogenetic relationships of its members. A group that contains its most recent common ancestor and *all* its descendants is said to be *monophyletic*. If some, but not all, descendants are contained, it is a *paraphyletic* group. If the most recent common ancestor is not in the group, it is said to be *polyphyletic*. Traditionally, classification has been based on the concept of a *grade*, i.e. a grouping determined on the basis of overall morphological similarity. Such groupings often do not reflect the precise genetic relationships between the species, and are frequently paraphyletic or polyphyletic groups. The alternative is a *clade*-based classification, determined on the basis of common genetic origin, or monophyly. Because both morphological similarity and genetic relatedness between species are such primary concerns, both grades and clades remain important for taxonomy (Cronquist, 1987; Sokal, 1985). On the basis of the computer simulations mentioned in Section 1.3, Sepkoski and Kendrick (1993) found that, for incomplete data, polyphyletic groups may be just as useful ‘systematically’ as are clades (monophyletic groups). The species simulations in this book employ both techniques (see Section 3.1).

The debate between these two approaches has led to a degree of confusion when discussing hominoid relationships, in particular with regard to the question of 'what is a hominid?' According to traditional hominoid classification, as discussed in Section 1.1, a hominid is any species more closely related to humans than to any other living hominoid, and the great apes are grouped as the pongids. Given what is now known about the closeness of the genetic relationships between humans and African apes, to the exclusion of orangutans, this grouping is no longer workable. To resolve this problem, two solutions have been proposed: either elevate chimpanzees, gorillas and orangutans each to their own family, or group gorillas, chimpanzees and humans as hominids (Gee, 2001). There is not yet consensus on which of these is the better solution, but the second suggestion appears to be gaining acceptance, with the consequence that all species formerly known as hominids are now placed in the tribe *Hominini*, and are thus known as *hominins*. For the purposes of this book, and in the absence of a consensus solution, I will continue to employ the traditional nomenclature, trusting that no ambiguity will result.

2.1 Hominoids

Figure 2.1 gives an overview of the current understanding of hominoid phylogeny (Goodman *et al*., 1998; Jones *et al*., 1991; Pilbeam, 1996; Shoshani *et al*., 1996; Begun, 2002). The Oligocene genera are considered by some to be ape ancestors, but are most likely primitive catarrhines (i.e. also ancestral to the Old World monkeys). Our understanding of hominoid fossil genera and species remains quite fluid owing to the scarcity of fossils, identification problems because of the substantial sexual dimorphism in many middle Miocene hominoids, and the many other difficulties associated with fossil studies.

Among the more notable features of Figure 2.1 is the large number of genera in the past. It is believed that between 22 and 16 million years ago (Ma) Africa was home to at least 14 hominoid genera.[1] However, by 14 million years ago, despite there still being several hominoid species, they were mainly found in Europe and Asia. This raises an interesting migration question for human origins: did the human–chimpanzee–gorilla common ancestor evolve in Africa, or did it evolve out of Africa and then return?

Hominoid migrations between Africa and Asia have been the subject of much recent research, with Stewart and Disotell (1997) attempting to tie

[1] Begun (2003) suggests up to 100 species between 22 Ma and 5 Ma.

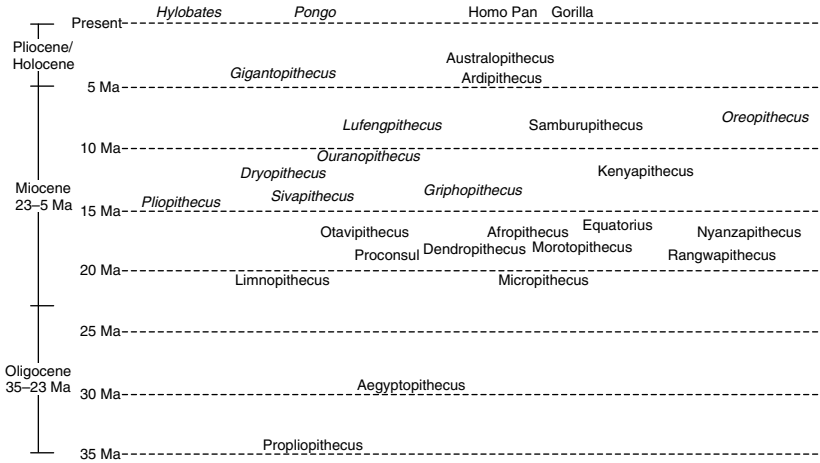


Figure 2.1. An illustration showing many of the known hominoid genera from the past 35 million years (Ma). Asian and European genera are in italics; the horizontal placement gives an approximate indication of morphological similarity, although no connections are drawn.

together various pieces of fossil, molecular and biogeographical evidence. Of particular importance are the global sea-level changes between 20 and 10 Ma that led to periods of connection and isolation between Africa and Eurasia, as well as various climate changes and their impact on species mobility. Begun (2003) provides a detailed discussion, describing how falling sea levels allowed the passage of African Miocene apes into Europe between 16 and 17 Ma, and thence to western Europe and the Far East. During a subsequent period of Eurasian isolation between 13.5 and 8 Ma, these apes evolved into a number of new forms, while there is a corresponding absence of African forms from *Kenyapithecus*, last in Africa around 12 Ma, to *Samburupithecus*, which appears in Africa around 9.5 Ma. Drastic climate change in the late Miocene led to many extinctions amongst the Eurasian forms, with only *Sivapithecus* surviving in Southeast Asia and *Dryopithecus* surviving by moving into the African tropics.

Some of the more important and better-understood genera are described below. *Proconsul* is considered by some to be a common ancestor for hominoids, but may instead be a sister taxon (Jones *et al.*, 1991; Pilbeam, 1996). Living from 23 to 14 Ma, it was a sexually dimorphic, tailless, arboreal quadruped with a diet of fruit and a body mass of up to 50 kg. Fossils have been found in East Africa, and no suspensory abilities are indicated. *Morotopithecus* (Gebo *et al.*, 1997) is a more gorilla-like genus from Uganda, also of body

mass up to 50 kg, but with some suspensory ability. *Equatorius* (Ward *et al*., 1999) is a controversial recent East African discovery that appears to be a primitive *Kenyapithecus*. Of body mass up to 30 kg, *Equatorius* occupied a dry, open woodland environment. *Sivapithecus* (Jones *et al*., 1991; Pilbeam, 1996), probably the best-known Asian hominoid fossil species, was arboreal and sexually dimorphic, weighed up to 90 kg and lived in a temperate woodland environment in South Asia between 13 and 7 Ma. Although the animal was not suspensory, the skull shows affinities with that of the orangutan. *Oreopithecus* has been receiving a great deal of attention recently, because of suggestions that it was bipedal (Gee, 2001; Köhler and Moyà-Solà, 1997; Rook *et al*., 1999). Fossils have been found in west and central Europe, but there is no suggestion of a direct link to any living hominoid.

The range of size variation in extinct hominoids is seen in the comparison of *Gigantopithecus*, a giant (up to 300 kg) Asian hominoid most closely related to orangutans, and *Micropithecus*, a 5 kg gibbon-like East African contemporary of *Proconsul*.

2.2 Hominids

Figure 2.2 gives an overview of the current understanding of hominid phylogeny, showing all known species and their corresponding time ranges. The evolution of hominids was thought to have followed a simple, linear path involving only australopithecines (robust and gracile) and humans but, as Figure 2.2 shows, this picture has become somewhat more complicated. Over the past decade, the discovery of *Australopithecus anamensis* (Leakey *et al*., 1995), *Australopithecus garhi* (Asfaw *et al*., 1999) and other species has expanded the genus *Australopithecus*, but other, even more recent finds have resulted in quite an increase in the number of genera. The additions of *Orrorin* (Senut *et al*., 2001), *Ardipithecus* (Haile-Selassie, 2001; White *et al*., 1994), *Kenyanthropus* (Leakey *et al*., 2001) and *Sahelanthropus* (Brunet *et al*., 2002) each result in a new genus, as well as extending the hominid fossil record back to 7 Ma: at or near the believed human–chimpanzee divergence time. (Some researchers are even proposing a reclassification of *Australopithecus afarensis* as *Praeanthropus africanus*, thus introducing a possible eighth hominid genus (Wood and Collard, 1999).

Fossils of the earliest species indicated in Figure 2.2 are exclusively African, with *Sahelanthropus tchadensis* from Chad (north-central Africa) and *Orrorin tugenensis* and *Ardipithecus ramidus* from East Africa. Species belonging to the genera *Australopithecus* and *Paranthropus* are also exclusively African.

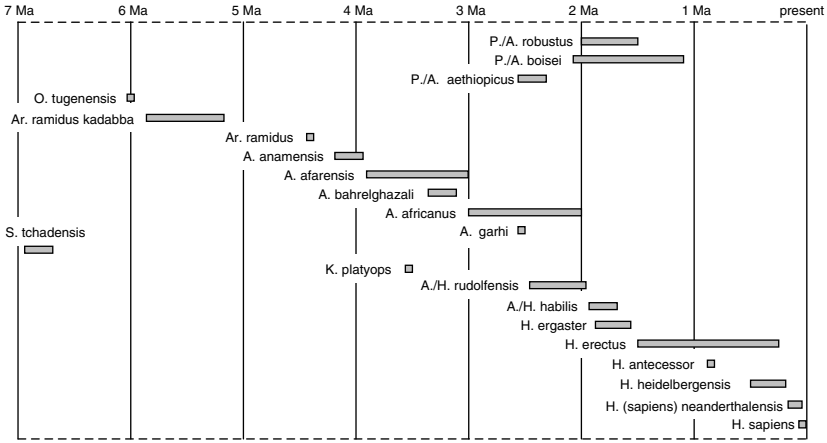


Figure 2.2. Illustration of the current understanding of hominid species. Genus abbreviations: S., *Sahelanthropus*; O., *Orrorin*; Ar., *Ardipithecus*; A., *Australopithecus*; K., *Kenyanthropus*; P., *Paranthropus*; H., *Homo*. Where the genus is in dispute, both possibilities are indicated.

Among the gracile species, *A. afarensis*, *A. garhi* and *A. anamensis* are all from East Africa, whereas the fossils of *A. bahrelghazali* were found in Chad and those of *A. africanus* in southern Africa (Wood and Richmond, 2000). Similarly there are regional variants of the robust species, with *P. boisei* and *P. aethiopicus* from East Africa, and *P. robustus* from southern Africa. The regional variation between these species is more pronounced than is the case for the gracile species (Jones *et al.*, 1991).

Both the eastern and southern African sites indicate a wooded habitat with open-country or grass-dependent fauna. Around 5 Ma climate shift in central and eastern Africa led to the formation of the Sahara desert, and there is also geological evidence of a shift to cooler, drier, more variable climates after 2.5 Ma (Behrensmeyer *et al.*, 1997). The period from 2.5 to 2.3 Ma has many extinctions and appearances in the vertebrate fossil record, the result of interactions between food resource and climate (Jones *et al.*, 1991).

The species *Homo habilis*, traditionally placed as the first member of the genus *Homo* and associated with the advent of tools around 2.4 Ma, has undergone substantial reassessment, and is still the subject of much debate. Firstly, the earlier fossils are now considered by many to belong to a separate species, *Homo rudolfensis*, and there is even the suggestion that both these species should be transferred to the genus *Australopithecus* (Wood and Collard, 1999; Wood and Richmond, 2000).

Homo erectus is the first hominid species for which there is undoubted evidence for spread beyond Africa, with fossils found in several Asian locations, and possibly in southern Europe. Early *erectus* fossils in Africa are now usually assigned to a separate species, *H. ergaster*. The subsequent migration of *Homo* species to Europe is apparent, with *H. heidelbergensis* found in each of Africa, Asia and Europe, and more restricted ranges for *H. antecessor* (Europe only), *H. neanderthalensis* (Europe and western Asia only) and early *H. sapiens* (Africa and western Asia).

All species shown in Figure 2.2 were (are) bipedal, and this has always been seen as an indication of membership of the hominid family. However, because of the situation with *Oreopithecus* described in the previous section, plus what is being learnt from the recent discoveries of very early hominids, this basic assumption is now being seriously questioned (Gee, 2001). Also, because the very early hominid species *Orrorin tugenensis* and *Ardipithecus ramidus kadabba* (Haile-Selassie, 2001) apparently occupied a woodland habitat, the traditional belief that the absence of chimpanzee fossils is due to their occupying forested environments unsuited to fossil preservation no longer has much support. This, combined with the question mark over the exclusiveness of bipedality to the human line, is putting far more pressure on the identification of some of these earlier fossils as human ancestors.

The exact impact on the above picture of the extremely recent discovery of the even older *Sahelanthropus tchadensis* (Brunet *et al*., 2002) is currently unclear. However, what is very clear is that the impact will be substantial (Wood, 2002). The known features of this 6–7 million year old species pose a significant challenge: for example, the braincase is very similar to that of a chimpanzee, whereas the canines and brow ridge are more like those of species of the genus *Homo*, and therefore do not appear elsewhere in the fossil record until several million years later.

Quite a different picture has been presented recently by Arnason *et al*. (2000), who, on the basis of molecular and fossil evidence, constructed a hominoid phylogeny substantially different from the current consensus. They applied a variety of phylogenetic methods, including parsimony, maximum likelihood and distance, to mtDNA sequences from primates and many mammals, chosen on the basis of the quality of their fossil records. Interpreting their results in a palaeontological context, they estimate the human–chimpanzee divergence time to be between 10.5 and 13 Ma: much further back than the research discussed above and in Chapter 1 indicates. Similarly, other important primate divergences are pushed back in time: the hominoid–Old World monkey split is placed at 52 Ma (cf. the commonly quoted value of between 25 and 30 Ma), and the New World monkey divergence is placed at 70 Ma (cf. the commonly quoted value of between 30 and 40 Ma). They

go on to consider the dispersal of ancient primates, relating their timings to continental movements.

These much earlier dates also impact on modern human origins, and Arnason *et al*. (2000) follow Senut *et al*. (2001) and present a phylogeny and timing differing substantially from that implied by Figures 2.1 and 2.2. In particular, *Ardipithecus ramidus* is placed on the chimpanzee line whereas *Orrorin tugenensis* and *Praeanthropus africanus* (*Australopithecus afarensis*) are placed on the human line, but only after a human–*Australopithecus* split at about 6 Ma. Two hominoid genera, *Samburupithecus* and *Ouranopithecus*, are placed on the *Australopithecus*–human lineage before this split, but after the chimpanzee lineage has diverged. Finally, they postulate a bottleneck in the modern human population at around 400 000 years ago, resulting from reduced genetic exchange owing to a hypothesised change in karyotype from $2n = 48$ to the current $2n = 46$.

Clearly there are issues concerning the association of fossils with extant lineages, resulting from non-hereditary characters and their effect on the determination of accurate phylogenetic relationships, and the confounding effects of extinction patterns and migrations. The first set of models described in this volume are designed to examine precisely these issues. The models are quite general in nature, but most runs will be designed with the particular context of hominoid and human origins in mind.

3 *Simulation design*

The species simulation has two distinct parts. The first involves simulating the evolution of species and is described below. The second involves taking the evolved species, randomly designating some of them as fossils, and attempting a reconstruction of the phylogeny by using current species and fossils (Section 3.1.1) as well as a reconstruction based on Wagner distances (Farris, 1970; Wiley *et al*., 1991) that uses current species only (Section 3.1.2).

Each simulation starts with a single ancestral species that has a particular suite of morphological characters. This species is then evolved in discrete steps, each step corresponding to a single *species generation*, and the nature of the available changes at each step is controlled by various parameters as described below. Each species generation is referred to simply as a *generation* in the following, but this should not be confused with generations of individual members of the species: this simulation has no concept of individual members of a species. Similarly, the immediate ancestor of a given species is termed its *parent* species, and the immediate descendants are called *child* species.

In each iteration, the evolved species are subjected to the effects of mutation, migration, extinction and possibly interbreeding. The degree and manner of influence from each of these effects is controlled via a number of user-specifiable parameters.

Modelling the complex interplay between hereditary and non-hereditary effects on morphological characters, described in Section 1.1, is an important aspect of the species simulation. As shown by Oxnard (2000), a whole-organism morphometric average reinforces the phylogenetic information while minimising functional parallels and convergences in local anatomical units, because the former is additive over different anatomical units, but the latter is independent across anatomical units. In the simulation, the suite of morphological characters is modelled as a fixed-length binary string, with some of the characters hereditary (so their states are passed on directly from the parent species), and the remaining characters non-hereditary. Each hereditary character is completely independent of the others, but the non-hereditary characters are modelled as a set, representing a particular environment–lifestyle combination. Rather than modelling the change in individual non-hereditary

characters, the simulation models a change in this environment, with consequent change in the non-hereditary characters. This process is controlled by three parameters: the number of non-hereditary characters; the number of distinct adaptive states these characters represent; and the probability of change.

For the initial species, all hereditary characters are arbitrarily set to zero, and no information on whether any particular character is hereditary or not is provided to the reconstruction algorithms.

Mutation is modelled by allowing a small number (possibly zero) of the morphological characters of a child species to differ from those of its parent. The immediate descendants of any given species must all differ from one another, but any one of them may be identical to the parent species. The nature of the mutation model is controlled by the following parameters. Firstly, there is the chance of any mutation occurring in a single species generation. Next is the number of characters affected, and this can represent either the maximum number of characters affected whenever there is a mutation, or the exact number of characters affected. This enables, for instance, simulation of punctuated equilibrium, by setting a low chance of mutation but a large number of characters affected whenever mutation occurs. Finally, a penalty can be applied to reversals (i.e. changes in binary state from 1 to 0), which when set to 0% allows reversals equally with forward mutations, when set to 100% totally excludes reversals, and for values in between reduces the likelihood of reversals in proportion.

Migration is modelled by giving each species a certain probability of migrating to any one of three other *continents*, labelled *A, B, C* and *D*. The likelihood of a migration is determined by a migration function that specifies a matrix of migration probabilities, where each value indicates the chance of migration between a particular pair of continents. This matrix is not necessarily symmetric, i.e. the probability of a migration from continent *A* to continent *B* need not be the same as that for a migration from continent *B* to continent *A*. A pair of generation times are specified for each continent such that migrations to this continent are only allowed between these times. It is also possible to specify that initial species should be placed on each continent, and that these should be either closely or distantly related.

The likelihood of any given species becoming extinct is determined by an extinction function that specifies an extinction rate as a function of generation, as well as two rates that override the specified values should the number of species go above a maximum or below a minimum value. Selective advantage can further alter the extinction probability for any given species. A default advantage value between 0 and 10 is assigned to each continent, and each

species has an associated value. A migrant will carry its value to its new location; if different from the default, its descendants will move closer to the default value by one each generation. Whenever a species has an advantage value greater than the default value for its continent, its likelihood of extinction is decreased accordingly. Correspondingly, extinction is more likely when the advantage is lower than the default.

If merging of lineages is being modelled, the simulation units can no longer be considered species and are more appropriately considered as either subspecies or interbreeding groups. The degree of interbreeding can be controlled by specifying the probability of merging given that two subspecies differ by less than some specified number of characters and have a common ancestor no further back than a specified number of generations. Alternatively, the desired number of lineage merges per species generation may be specified. It is also possible to restrict the range of species generations over which merging may occur. It is important to remember that even under these conditions it is species that are being simulated, but the one-to-one mapping between simulation units and species has been broken. Rather, one or more simulation units combine to form a species, as can be determined by following the simulated lineages.

A further important feature of the simulation is the ability to broadly constrain the number of species per generation, i.e. the *diversity profile*, so that only distributions with particular features are included when multiple runs are made and averaged. For example, we may wish to study only distributions with larger numbers of ancestral species, but few present-day species, approximately matching the distribution of fossil and living hominoids. The simulation allows specification of any or all of the following:

- minimum and maximum final population;
- minimum and maximum population for any generation;
- start and end generations for the application of the previous limit;
- whether the limits are applied across all continents, or to specific ones independently.

It is also possible to ensure merging in the history of a surviving lineage if desired.

The maximum number of child species that may arise from a single parent species in any generation can also be specified. This setting, along with the mutation rate, acts to provide a time scale for the simulations. The settings used for the species simulations presented in the following chapters are such that each simulation generation can be considered as corresponding to a period of roughly one million years.

3.1 Phylogenetic reconstruction

After generating a tree of related species, a *species preservation* rate (fixed or varying linearly with generation) that specifies the likelihood that any particular species is present in the fossil record is applied, leaving a number of present-day species plus a random smattering of fossil species. This rate will be referred to simply as the fossilisation rate, with the understanding that it applies to species rather than individuals. The problem is then to use these data to reconstruct the phylogeny and migration history from knowledge of the living and fossil forms alone, and then compare the results against the known true phylogeny from the simulation.

Two techniques of reconstruction are employed. The first technique involves matching both existing species and fossils to the most closely (morphologically) related earlier fossil (see Section 3.1.1). The other involves a reconstruction based on the discrete morphological characters of the existing species only (see Section 3.1.2).

3.1.1 Reconstructing by using fossils

The information available to the fossil reconstruction consists of all current species and a number of earlier species known from fossils. The morphological characters and ages of the fossils are known, but it is not known (to

Algorithm 1 Reconstruction algorithm

1. *Choose a current species or fossil.*
2. *Calculate the character difference between it and all earlier fossils.*
3. *Choose the most recent fossil, on the same continent if possible, with the least difference.*
4. *See whether this fossil is a possible ancestor by comparing the degree of character difference with the generation difference, by using an estimate of the divergence for any generation. Since non-hereditary adaptation can lead to quite substantial variation between a parent and a child species, the estimated divergence per generation is generally somewhat larger than the specified mutation rate.*
5. *If the fossil was found to be a possible ancestor, make a connection.*
6. *Repeat from step 1 until all fossils have been considered.*

the reconstruction) which of the characters are hereditary. The reconstruction algorithm (shown as Algorithm 1) traces backwards in time, matching species and fossils as well as possible, and produces a hypothetical phylogeny, which is then analysed and compared with the known true phylogeny. Once the reconstruction is completed, clades are identified by using Algorithm 2. Acceptance of a clade in step 5 of Algorithm 2 is determined by considering three constraints: the minimum number of species, the minimum total size of the clade and the maximum total size, where clade size is a function of the number of species included as well as their durations. However, each of these constraints is user-configurable.

Algorithm 2 Clade identification algorithm

1. *Rank all fossils and current species based on the distance to their four previous ancestors. (This is an arbitrarily chosen measure, but one that captures well implicit groupings in the fossil connections for the usual size of simulation run.)*
2. *Choose a fossil with the least distance according to this measure.*
3. *Find its immediate parent.*
4. *Construct the group comprising this parent and all its ancestors.*
5. *Check whether this satisfies the definition of a clade.*
6. *If the group is too large, go to step 10.*
7. *If the group is too small, find the parent's parent, if it exists, and go to step 4. If the parent's parent does not exist, go to step 10.*
8. *The group satisfies the definition of a clade, so label all group members as members of this clade, and continue.*
9. *Remove any newly labelled fossils from the set of unclassified fossils.*
10. *Choose the next remaining fossil from the original ranking, and repeat from step 3. Finish when all fossils have been considered.*

3.1.2 Wagner reconstruction

An alternative approach is to use the characters of the existing species and construct a phylogeny on that basis alone. Nei (1991) discusses various methods; given the idealised situation of the simulation, a distance method is fastest and sufficiently accurate (Farris, 1970; Wiley *et al.*, 1991).

The molecular clock hypothesis is perfectly true for the simulation (at least for the mutation of hereditary characters), so the degree of character difference is directly related to the time of divergence and there is no gain in employing a more complicated and time-consuming phylogenetic algorithm such as maximum likelihood or parsimony (Felsenstein, 1982). Furthermore, Ruvolo (1997) analysed all human DNA data sets available at the time, and only in one case (out of 14) did the distance method result in a phylogeny different from that obtained by a parsimony method.

A *phenetic* reconstruction involves grouping species based on overall similarity without distinguishing between shared primitive and shared derived states, and produces a *phenogram*. This is essentially the process used in the fossil reconstruction described above (Section 3.1.1). Cladistics, as discussed in the introduction, involves the identification of shared derived character states, and produces a *cladogram*, i.e. a diagram that aims to indicate actual rather than apparent relationships. Distance methods do not provide a cladistic reconstruction automatically because of their failure to differentiate between primitive and derived characters. However, specification of an outgroup does allow this identification, particularly in the context of the simulation, as the following calculation shows.

Considering a single lineage of species with n characters, evolving over $2m+1$ generations with an average mutation rate of one character reversing every two generations and no reversal penalty, the chance of any particular character reversing k times is given by

$$P(k) = \left(\frac{1}{n}\right)^{k} \left(\frac{n-1}{n}\right)^{m-k} \binom{m}{k}.$$

Assuming an outgroup with a character vector consisting entirely of zeros (as is the case for the simulation), binary character states with value 1 must be derived. A more interesting question is: what is the likelihood that a character state of 0 in an evolved species is primitive? From the above result, the fraction of sites for the species at the end of a lineage that will have had no reversals is given by

$$P(0) + P(1) = \left(\frac{n-1}{n}\right)^{m-1} \left(\frac{n-1+m}{n}\right).$$

The simulations presented in the next chapter mostly associate a vector of 24 hereditary characters with each taxon, mutating either not at all, or by a single character each generation, over 25 generations. Substituting $k = 24$ and $m = 12$ (25 generations implies 24 evolutionary steps and an associated

expectation of 12 mutations) gives $P(0) + P(1) = 91\%$. Therefore, for any species at the end of a lineage, we would expect 13 characters to have retained their primitive state of 0, and of the 11 that have changed, one of them will have reversed. So, on average, 13 of the 14 characters with value 0 (i.e. 93%) are truly primitive: crucial to the identification of the above method as a cladistic method.

To apply the Wagner distance method, each taxon is represented as a vector of n binary characters

$$T = \{T_n\} = t_1 t_2 \ldots t_n,$$

and the distance between any two taxa is given by

$$d(T_1, T_2) = \sum_{i=1}^{n} \chi(T_1, T_2, i)$$

where χ is a function that counts the number of sites where T_1 and T_2 differ, defined as

$$\chi(T_1, T_2, i) = \begin{cases} 1 & \text{if } t_{1i} \neq t_{2i} \\ 0 & \text{otherwise.} \end{cases}$$

Algorithm 3 Wagner algorithm

1. *Specify an ancestor or outgroup; call this taxon O.*
2. *Calculate and store each* $d_n = d(T_n, O)$.
3. *Choose the taxon with the smallest* d_n.
4. *Create an interval for this taxon by joining it with O.*
5. *Choose the taxon with the next smallest* d_n.
6. *From the phylogeny so far constructed, find the interval with the smallest distance for this taxon (using the equation for interval–taxon distance given in the main text).*
7. *Attach the chosen taxon to this interval by constructing an ancestor, with character vector the median of the taxon and the two taxa that define the chosen interval.*
8. *Repeat from step 5 for all remaining taxa.*
9. *Optionally optimise the hypothetical taxa by using the* ACCTRAN *or* DELTRAN *algorithms described by Wiley* et al. *(1991).*

Also needed is the distance between a taxon and an interval Int(T), defined by a taxon T and its immediate ancestor Anc(T):

$$d(T_1, \mathrm{Int}(T_2)) = \frac{1}{2}[d(T_1, T_2) + d(T_1, \mathrm{Anc}(T_2)) - d(T_2, \mathrm{Anc}(T_2))].$$

Given these definitions, the calculation of the phylogeny for a set of taxa $\{T_n\}$ follows the steps shown in Algorithm 3. After constructing the phylogeny, fossils can be matched to the hypothetical taxa where possible, but this is a difficult process because of the lack of an inherent time scale in the Wagner reconstruction process. However, clades identified from the fossil reconstruction and containing current species are compared with the Wagner reconstruction, and the extent of agreement measured.

The *ACCTRAN* and *DELTRAN* algorithms optimise Wagner trees, leaving the topology unchanged while optimising the characters of the hypothetical taxa in one of two ways. *ACCTRAN* accelerates the transformation of characters, making any transformation occur at the earliest possible taxon, and thus favours parallelisms over reversals. *DELTRAN* delays character transformations to the latest possible taxon, and thus favours reversals over parallelisms (Wiley *et al.*, 1991).

A time scale is imposed on the Wagner tree by finding the most recent common ancestor (which always exists), calculating the average length of the branches leading to the common ancestor and converting this number into a generation by using the expected number of changes per generation. The expected number of changes per generation is itself difficult to determine. The simulation does so by averaging the number of derived character states across all current species and dividing by one less than the total number of generations (one less because n generations implies $n-1$ iterations with mutation). The accuracy of this estimate is limited owing to character reversals, because in such cases the primitive and derived states are indistinguishable. Correcting for this would imply faster changes of character state, and thus act to bring the Wagner common ancestor closer to the present time. However, the actual effect is relatively small (as discussed above), and already the simulation has a substantial advantage in that the total number of generations is known exactly. In real life, of course, this is not known and must be estimated from the fossil record.

3.2 Example simulation and reconstruction

Figure 3.1 shows a simple example simulation of evolution over eight generations, with eight characters per species, leading to four current species.

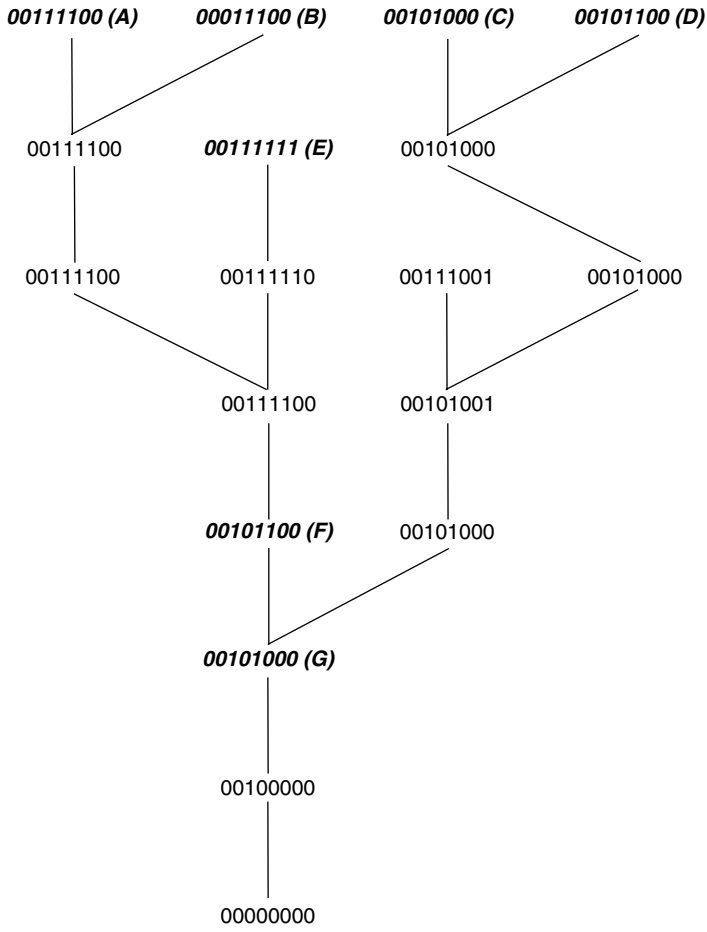


Figure 3.1. A sample simulated phylogeny, with at most a single character mutation each generation. Fossil species are shown in bold italics, followed by a single-letter label.

This sample is used below to illustrate the workings of both reconstruction algorithms.

After applying the fossil reconstruction algorithm, the closest fossil to each of the current species and fossils can be identified. The results are shown in Table 3.1. The fossil labelled F is closest to three of the four current species. The other current species, C, as well as F itself, are determined to be most closely related to fossil G, the earliest fossil in the simulation. In each

Table 3.1. *Characters, closest fossil and number of differences for each current species and fossil from the example simulation shown in Figure 3.1*

Species	Generation	Characters	Closest fossil	Differences
A	8	0011 1100	*F*	1
B	8	0001 1100	*F*	2
C	8	0010 1000	*G*	0
D	8	0010 1100	*F*	0
E	7	0011 1111	*F*	3
F	4	0010 1100	*G*	1
G	3	0010 1000	—	—

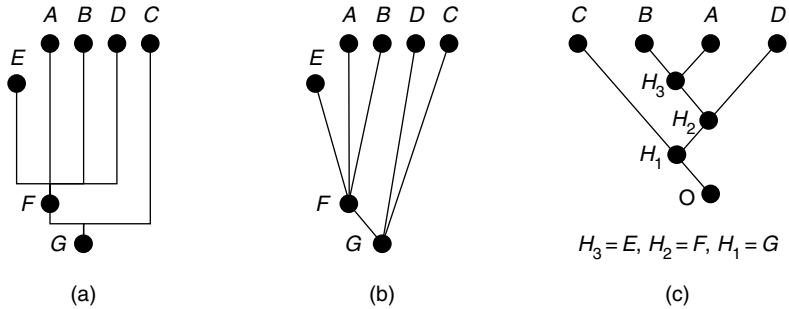


Figure 3.2. Reconstruction for the sample fossils (a), true fossil connections (b) and Wagner reconstruction with hypothesised ancestor taxa (c). The topologies of the two reconstructions are very similar, but both differ from the true phylogeny.

case the closest fossil is a possible ancestor, so the reconstruction makes the connections as shown graphically in Figure 3.2a.

To apply the Wagner reconstruction algorithm to the sample simulation, we first note that characters 1, 2, 7 and 8 have identical states for all current species and thus give no information and can be ignored. The inputs are therefore:

$$A = 00111100 = 1111$$

$$B = 00011100 = 0111$$

$$C = 00101000 = 1010$$

$$D = 00101100 = 1011$$

$$O = 00000000 = 0000,$$

where A, B, C and D are the current species from above, and O is the outgroup.

The distances from the outgroup for each current species are

$$\begin{aligned} d(A, O) &= 4 \\ d(B, O) &= 3 \\ d(C, O) &= 2 \\ d(D, O) &= 3, \end{aligned}$$

so the first step is to attach C, the closest species, to the outgroup O. Next closest is a tie between B and D, so choosing D arbitrarily, and attaching it to the interval OC, gives $H_1 = 1010$. Now taking B (equal second closest with D), its distance from all intervals is calculated:

$$\begin{aligned} d(B, AH_1) &= 3 \\ d(B, DH_1) &= 2 \\ d(B, H_1O) &= 2. \end{aligned}$$

Again there is a tie,[1] so the interval DH_1 is arbitrarily chosen and B attached to give $H_2 = 1011$. Only A remains to be added, so again after calculating the distances for all current intervals:

$$\begin{aligned} d(A, DH_2) &= 1 \\ d(A, BH_2) &= 0 \\ d(A, H_1H_2) &= 1 \\ d(A, CH_1) &= 2 \\ d(A, H_1O) &= 2, \end{aligned}$$

A is attached to BH_2 to obtain $H_3 = 1111$.

The final Wagner tree is shown in Figure 3.2c. In this case, there is a lucky matching of hypothetical ancestor character states with those of true fossil species. (This is completely true for species F and G, but species E has values of 1 at positions 7 and 8 that are ignored in the Wagner reconstruction because they don't help to differentiate current species.)

[1] Obviously there are fewer ties when more variation is allowed than in this restricted example.

As discussed above, because the outgroup has all states 0, character state 1 is *always* a derived state, whereas character state 0 is only derived if it is the result of a reversal. Therefore, a derived 0 is much less likely than a primitive 0 because it requires an initial random change from 0 to 1, and then the same character to randomly change again. In the above example, for instance, although there are only eight characters, only two reversals occur out of a total of 11 state changes. So, considering the four living species in the example shown in Figure 3.1, of the three sites where they have differences, species *A*, *C* and *D* share a derived state of 1 for character 3, species *A* and *B* share a derived state of 1 for character 4, and species *A*, *B* and *D* share a derived state of 1 for character 6. This indicates that *A*, *B* and *D* most probably form a clade, with *A* and *B* most closely related, and thus diverging most recently, with *C* splitting off first of all. This is indeed the tree found by using the Wagner algorithm and shown in Figure 3.2c. The alternative explanation of *A*, *C* and *D* forming a clade cannot adequately explain the derived character state shared by *A* and *B*.

As can be seen from Figure 3.2, there is substantial agreement in topology between the two reconstructions. However, neither manages to recognise the close relationship between species *C* and *D* seen in Figure 3.1. This is partly due to the restricted character space producing a number of accidental similarities, and partly due to the lack of fossils on the lineage leading to *C* and *D*.

Perhaps the most important result obtained from the reconstructions is the time of the most recent common ancestor of all current species. In the above example, not only is species *G* this common ancestor, it is also a fossil and correctly identified in the fossil reconstruction. The Wagner reconstruction does not involve the fossil species but rather, as shown above, constructs hypothetical ancestral taxa. In the example, the common ancestor is the hypothetical taxon labelled H_1, and has character states identical to those of species *G*. The time depth of taxon H_1 is not directly apparent from the diagram but must be calculated, as explained above, by looking at the average number of state changes along all paths from current species to H_1 and then using an estimate of the mutation rate. The path lengths for the Wagner reconstruction are

$$\begin{aligned} C \rightarrow H1 &= 0 \\ D \rightarrow H2 \rightarrow H1 &= 0+1=1 \\ A \rightarrow H3 \rightarrow H2 \rightarrow H1 &= 0+1+1=2 \\ B \rightarrow H3 \rightarrow H2 \rightarrow H1 &= 1+1+1=3 \end{aligned}$$

and so the average length is 1.5 mutations. The estimated mutation rate is found from the distances from the outgroup calculated earlier, averaged per species and per generation, to give $(4+3+2+3)/4/7 = 3/7$ characters per species per generation. These two values then combine to give an estimated generation of $8 - 1.5/(3/7) = 4.5$, which, after truncation, gives generation 4, i.e. one generation more recent than the true common ancestor.

3.3 Analysis and evaluation

The results of any simulation run are displayed graphically in a number of different ways, and also as a text summary that includes the subsequent analysis. For a single run, the graphical display shows all species (fossil and otherwise) and allows interactive querying of the true and reconstructed phylogenies. Examples of the display and available interactions are given in the following chapter. The text summary includes information about all parameter settings, individual fossil characters and migrations, and analysis of common ancestry, fossil connections, and clade identification. The precise information recorded is described in detail in Section 3.3.1. For an average run, individual species data are not available, but average diversity, accuracy, migration and clade identification information is given, with standard deviations where appropriate, as described in Section 3.3.2.

Also reported is the *seed* used to drive the random number generation within the simulation. Knowledge of this seed, plus all parameter settings, allows the exact reproduction of any simulation run.

3.3.1 Single-run output data

The single-run summary information can be grouped under six categories as follows.

1. Species and common ancestry:
   - the number of species, overall and each generation;
   - the number of new species each generation;
   - the number of extinctions each generation;
   - the average species duration;
   - the average pairwise character difference between current species;
   - the time and location of the true most recent common ancestor, the true most recent fossil common ancestor (if it exists), and the fossil reconstruction common ancestor (if it exists);

- the most recent common ancestor according to the Wagner reconstruction;
- each of the above values reported both overall, and for each continent.

2. Fossils and reconstruction accuracy:
   - the number of fossil species;
   - the percentage of fossil species connected to current species, both real and according to the reconstruction;
   - the number of fossils accurately reconstructed (i.e. correctly connected to their immediate fossil ancestor);
   - the number of current species accurately reconstructed.

3. Clade identification and analysis:
   - the number of clades, their total size, and number of current species included in each;
   - the true nature of each clade as either monophyletic, paraphyletic or polyphyletic (all clades are, of course, monophyletic according to the fossil reconstruction);
   - for clades that are actually paraphyletic, the percentage of true descendants included;
   - for clades that are actually polyphyletic, the percentage of true descendants included, and the percentage of included fossils that are actually not descendant;
   - the connection accuracy within each clade.

4. Analysis of clades containing current species:
   - whether the current species clade members are actually all the extant members of a single true clade (regardless of any paraphyly or polyphyly in the clade as a whole);
   - if the current species clade members are not monophyletic in the above sense, the percentage of current species found (with respect to the number of true current descendants of their most recent common ancestor);
   - the above two values calculated with respect to the Wagner reconstruction common ancestor rather than the true phylogeny.

5. Migrations:
   - a list of all migrations by generation, showing source and destination continent, and number of species;
   - a comparison of real, fossil record and reconstructed fossil migrations according to source and destination only (i.e. ignoring the time of migration);

- details of all migration times (real, start, end and average generation) according to the reconstruction, as well as the end times of migrations evident from the fossil record.

6. Detailed fossil information:
   - a list of all fossil species, grouped by clade, and including generation, continent, number of descendants, and characters.

3.3.2 Average-run output data

The average-run summary information is slightly different, and can be grouped under the following five categories.

1. Species and common ancestry:
   - the average number of species (total and per generation);
   - the average number of new species per generation;
   - the average number of extinctions per generation;
   - the average species duration;
   - the average pairwise character difference between current species;
   - the average time of the true most recent common ancestor, the true most recent fossil common ancestor (if it exists), the fossil reconstruction common ancestor (if it exists), and the Wagner reconstruction common ancestor;
   - the sample standard deviation associated with each of these values;
   - the distribution of locations for the common ancestors (except for the Wagner common ancestor since the Wagner reconstruction does not include location information);
   - each of the above values reported overall, and for each continent separately.
2. Fossils and reconstruction accuracy:
   - the average number of fossil species;
   - average percentages of fossils connected to current species, both real and according to the reconstruction;
   - the average number of fossils accurately connected to their immediate fossil ancestor in the reconstruction;
   - the average number of current species accurately reconstructed;
   - the sample standard deviation associated with each of these values.

3. Migrations:
   - the average numbers of migrations by source and destination continent;
   - average details of all migration times (real, start, end and average generation) according to the reconstruction, as well as end times of migrations evident from the fossil record;
   - the sample standard deviation associated with each of these values.
4. Clade identification:
   - number of clades, size, and number of current species, each averaged over all runs;
   - the total number of clades, their average size, and the number that included current species;
   - the percentage of identified clades that were actually monophyletic;
   - the percentage of identified clades that were actually paraphyletic, and the percentage of true ancestors included in these cases;
   - the percentage of identified clades that were actually polyphyletic, the percentage of true descendants included in these cases, as well as the percentage of included fossils that are actually not descendant.
5. Analysis of clades containing current species:
   - the percentage of times those current species were actually all extant members of a single true clade (regardless of any paraphyly or polyphyly in the clade as a whole);
   - if the current species clade members are not monophyletic in the above sense, the average percentage of current species found (with respect to the number of true current descendants of their most recent common ancestor);
   - the above two values calculated with respect to the Wagner reconstruction common ancestor rather than the true phylogeny, giving a measure of the correspondence between the two reconstructions. (It is important to take into account the current species clade accuracy measure for the fossil reconstruction when interpreting this value.)

4 *Running the simulation*

The results presented in this chapter give an overview of the features of the simulation while illustrating its operation. Starting with two very simple examples, the simulation of species and subsequent reconstruction from fossils and character information is shown. Next, with migration included, an example is given showing tracing of ancestry, and the three-dimensional display of intercontinental migrations. This example is also used to illustrate the automatic clade generation algorithm. The more advanced features of including non-linear mutation rates, non-hereditary characters and interbreeding are shown in the chapter's final examples.

In the discussion of the simulation results below and in subsequent chapters, the following abbreviations are employed:

CA, true most recent common ancestor;

WCA, most recent common ancestor according to the Wagner reconstruction;

FCA, most recent fossil common ancestor;

RCA, most recent fossil common ancestor according to the fossil reconstruction.

Note that 'most recent' is always implied by the abbreviation, although 'MR' is not added in each case.

4.1 A simple example

The two examples displayed in this section have been kept simple deliberately, in order to illustrate more clearly the fundamental aspects of the simulation, without the complicating effects of the many more advanced features available. Both simulations were run with the same parameters, so they also serve to emphasise the fact that the degree of variation between any two runs can be quite substantial.

Setting the character vector to a size of 32 independent states, with a 50% chance of one state changing in a single generation, the simulation

was run for 25 generations. The fossilisation rate varied linearly from an initial 10% to nearly 27% at generation 25. The extinction function was such that at generation 2 the extinction rate was 10%, increasing to 20% at generation 6, 30% at generation 10, and 40% at generation 13. (Note, that the extinction rate remains constant in between these generations, i.e. there is no interpolation.) If at any time the number of species dropped to 2 or fewer, the extinction rate was set to zero for that generation only. Conversely, if the number of species increased above 20 at any time, the extinction rate was set to 80% for that generation. There was no reversal penalty, no migration, and no interbreeding, and all characters were hereditary.

Figure 4.1 shows one run with these settings, with a small number of evolved species. Figure 4.1a shows the complete simulation, each circle representing an individual species, the filled circles being those species that are fossils or are members of the current generation, and thus accessible to morphological examination. The lines connect any particular species with its immediate ancestor. Figure 4.1b shows the same simulation, but only fossil and current species are displayed, and this time the connections show the link between any particular species and its most recent fossil ancestor. Figure 4.1c also shows fossils and current species only, but now the species are arranged and connected according to the reconstruction algorithm described in Section 3.1.1.[1] Heavier lines are used to indicate the grouping of species into clades.

Immediately obvious when comparing Figures 4.1b and 4.1c is the fact that a minor error in the reconstruction, involving the circled species, has a significant effect on the clade identification accuracy. The circled species is identified as ancestral to all the living species, and subsequently made the parent of the clade containing all living species, when in fact it is ancestral to none of them! It is certainly closely related to the true ancestral parent, but nevertheless lies on its own, now extinct, lineage. The most recent common ancestor (CA) of all living species can be seen from Figure 4.1a to occur at this generation (generation 20, numbered from an initial generation of 1), and to be a sister species to the circled species, yet despite (or perhaps because of) the high degree of fossilisation at this generation, the reconstruction algorithm fails. It is also interesting to note that despite the fact that over the last five generations the diversity is never greater than three species, two independent lineages have managed to persist. In this important respect, the reconstruction

[1] The location of the species in Figure 4.1c is different from that in Figure 4.1b because it reflects the reconstruction rather than the original simulation. *Specialist* allows both the reconstructed connections and the true fossil connections to be displayed with the species in the positions determined by either the original simulation or the reconstruction as required.

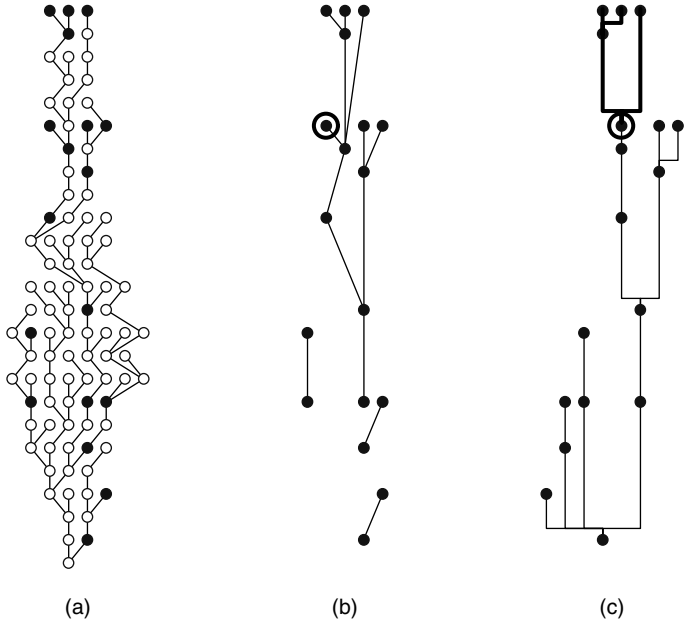


Figure 4.1. A simple simulation, resulting in few species, showing the full simulation results (a), fossil species only (b), and the reconstruction based on fossil matching (c). The fossil incorrectly identified in the reconstruction as the parent of the clade comprising all current species is highlighted with a circle.

is correct. The estimate of the time of the CA from the Wagner reconstruction is generation 18, so this method also correctly determines that there is a relatively deep split in the phylogeny of the living species. Of the 92 species generated, 15 have been randomly designated fossils, and of these only 5 lie on extant lineages (as can be easily verified from Figure 4.1b). In this example, the reconstruction places 7 of the 15 fossils on extant lineages. As will be seen as more results are presented, the reconstruction regularly overestimates the number of fossils on lineages leading to living species, often by a far greater degree than in this case.

Figure 4.2 shows another run with these same settings, this time producing a much larger number of evolved species, and showing a number of characteristics differing from those of Figure 4.1. The feature that stands out immediately is the growth in diversity over the first 13 generations, prior to a significant and sudden reduction. In fact, from this diversity maximum of 27 species, only one (in the dashed circle) has any descendants at all even five generations later! As a result of the mass extinction from generation

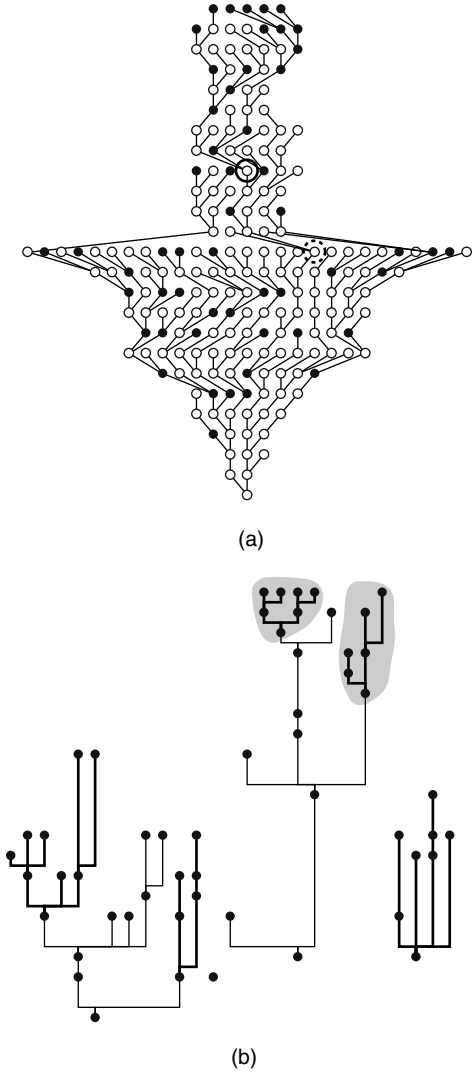


Figure 4.2. Another simple simulation, this time with many species, showing the full simulation results (a) and the reconstruction based on fossil matching (b). The species highlighted by a continuous circle is the most recent common ancestor of all living species, and the species enclosed by the dashed circle is the only species from the time of maximum diversity whose lineage has survived to the present time. The shading on the reconstruction highlights the two clades containing all current species (this time correctly identified).

13 to 14, even though there are 48 fossil species, only 10 of these fossils lie on lineages leading to living species. The CA of all living species is indicated in the figure by a continuous circle, and is at generation 17. Its parent fossil, two generations back, is the most recent fossil common ancestor (FCA) of all living species, and this fossil is correctly identified in the fossil reconstruction. The Wagner reconstruction again slightly overestimated the time required to generate the observed diversity of living species, hypothesising the CA at generation 15.

The clades indicated in Figure 4.2b show a high degree of accuracy; in fact the two clades involving living species (shaded in the figure) are 100% accurate. Of the five clades identified, four are truly monophyletic. The oldest clade, with six members and central in the figure, is actually polyphyletic, with two of the included species not actually ancestral to the parent species.

4.2 Migration

The addition of migration greatly complicates the situation described in the previous section. It leads to increased morphological diversity in any continent, with consequent difficulties in determining common ancestry and fossil connections. Figure 4.3 shows the results from a four-continent, 18 generation simulation, based on a fossilisation function and an extinction function that are the same as for the previous example. The migration settings are such that between generations 5 and 15 there is a 3% probability of migration between any two continents, with no migration possible at other times. In addition, there is no selective advantage for species from any particular continent.

Each continent is displayed separately; in the interests of clarity, connections are not drawn for migrant species. However, a migrant species is drawn as a (slightly rounded) square rather than as a circle, and in the colour (or shading) of its original continent. Obviously, the source species for continents *B*, *C* and *D* is always a migrant; in this example it is a migrant from continent *A* in all three cases. In addition, the option of indicating continuing species by a thicker line is employed.

One kind of interaction available with the simulation output is the display of all ancestors and descendants of any particular species. Clicking the mouse on any species causes all its ancestors and descendants, on any continent, to be drawn in red. Figure 4.3b shows only those species that are highlighted when the second-generation species on the left is clicked, and Figure 4.3c shows the highlighted species for the second-generation species on the right. This makes clear a number of aspects of the evolution that are not at all easily seen in Figure 4.3a. For example, the CA of all species is the original species on

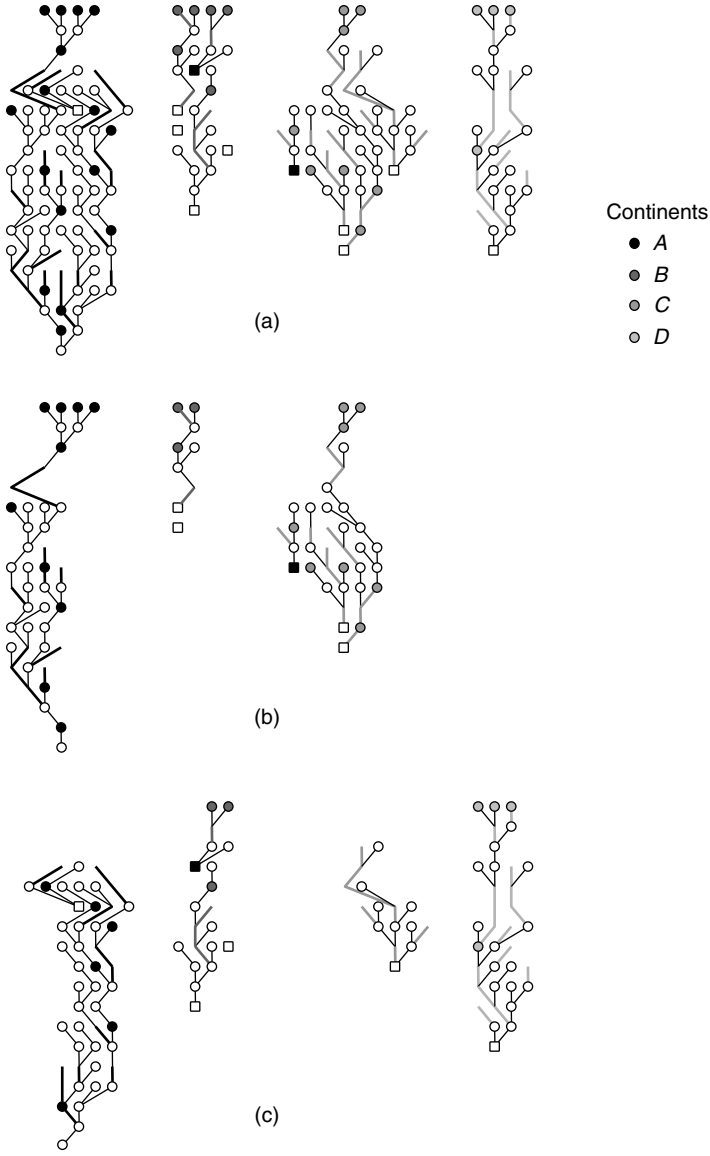


Figure 4.3. Full simulation involving migration between four continents; (b) and (c) show the two distinct lineages in isolation. In the simulation, lineages are highlighted by selecting a species, and holding the Alt key causes the relevant migrations to be shown as well. Migrant species are drawn as rounded squares rather than circles, and in the colour (or shading) of their source continent. Control-clicking a species at either end of a migration causes the connection to be explicitly shown. Continuing species are here indicated by a thicker line.

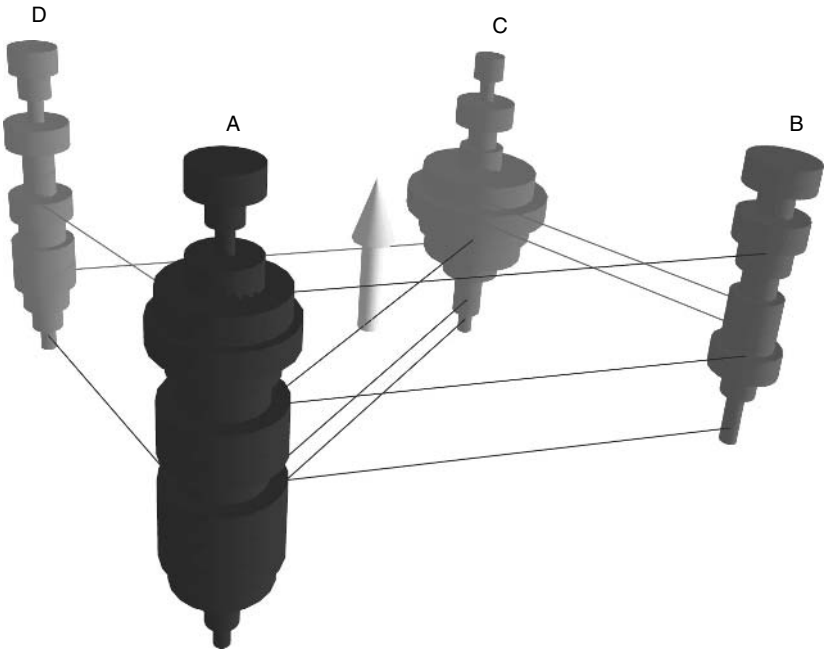


Figure 4.4. Another way to visualise continental diversity and migrations is with this 3D view, where the diameter of the component cylinders is proportional to the number of species that generation, and the migrations are shown by lines connecting source and destination, in the colour (or shading) of the source continent.

continent *A*, and all descendant species are approximately evenly split between being descendants of each of its child species. Continents *A*, *B* and *C* have all had many species descended from each of these original child species, but only on continent *B* are there living species from both lineages. Continent *D*, on the other hand, has only ever been populated by descendants of the second child species (the one on the right in the figure). The Wagner reconstruction is unable to correctly time the CA of the species on continent *B*, estimating generation 12 when the true value is generation 1. However, it does correctly determine that the generation of the CA of all species is well back in time, returning an estimate of generation 0.

Overall, 11 migrations have occurred: three from *A* to *B*, three from *A* to *C*, one from *A* to *D*, two from *C* to *B*, one from *D* to *A* and one from *D* to *C*. These can be most easily seen by using the 3D view shown in Figure 4.4. In this figure, each generation is represented by a disc, the relative size of which indicates the diversity, and migrations are shown by lines joining the

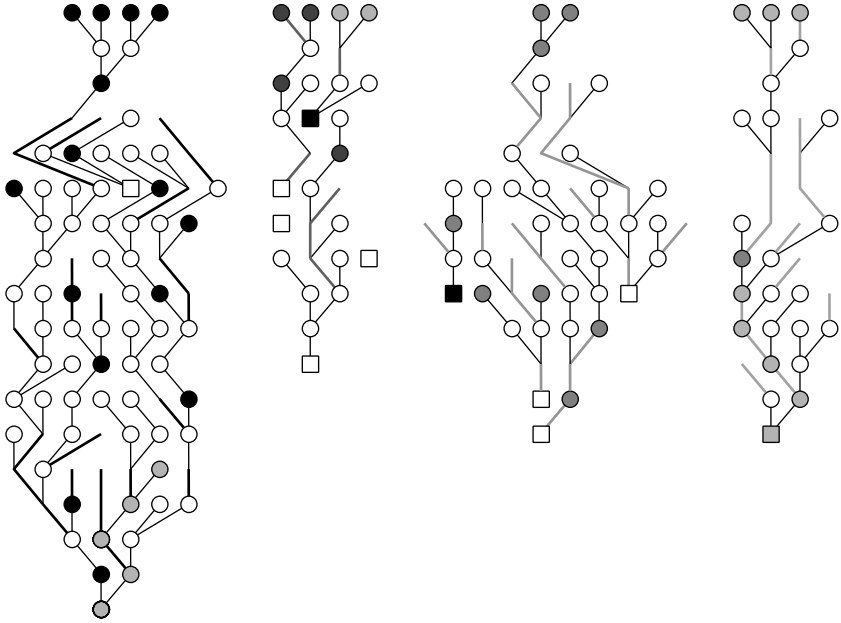


Figure 4.5. The simulation also displays common ancestors graphically. In this figure, all current species on continent *D* and two current species on continent *B* have been selected by double-clicking (these are drawn with red hatching in the simulation). The cross-hatched species in the earlier generations (drawn in blue in the online simulation) are the common ancestors of the selected ones.

source and destination continents, drawn in the colour corresponding to the source continent. It is a simple matter to match each line on the figure with one of the migrations listed above.

Similar to tracing ancestry, another online interaction available allows the visual determination of common ancestry. When a species is double-clicked, it is drawn in a red cross-hatched pattern as shown in Figure 4.5. (Fossil species are drawn with the inverse pattern.) Drawn with blue cross-hatches are then all species that are common ancestors of the set of currently selected (by double clicking and thus drawn in red) species. Figure 4.5 shows all common ancestors for the current species descended from the right-hand second-generation species of Figure 4.3. It is easy to verify that these two figures are providing a consistent picture.

The reconstruction for this simulation is shown in Figure 4.6a–c, each with a different set of clade identification parameters, and the true fossil connections are shown in Figure 4.6d. Figure 4.6a has a minimum number of

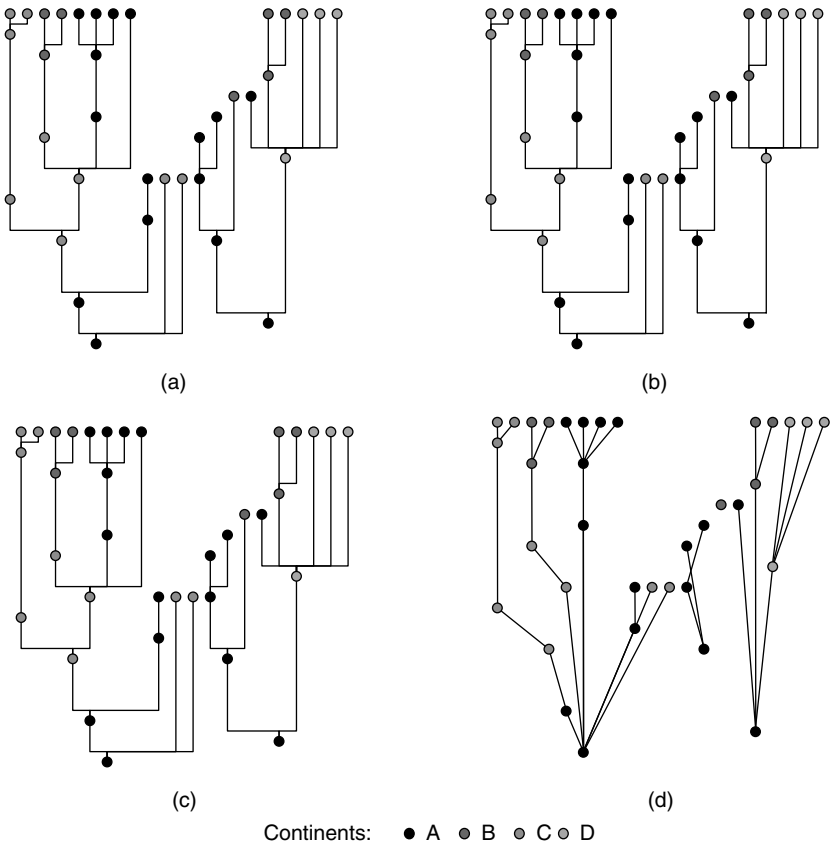


Figure 4.6. This figure shows the fossil reconstruction from the run displayed in the previous two figures, with three different settings for the cladistic classification: (a) minimum species/minimum size/maximum size set to 3/10/50 respectively, (b) 3/20/50 and (c) 3/50/100. The fourth figure (d) shows the true connections.

species of 3, a minimum size of 10 (size is determined by a combination of the number of species and the length of the included lineages, as discussed in Section 3.1.1), and a maximum size of 50. Five clades are identified, with living species spread across four of them and one living species unclassified. The depth of the clades implies that this is quite a reasonable division, although it is perhaps unsatisfactory to have the eight closely related living species in different clades. In Figure 4.6b, the minimum size is increased to 20 and there are now only two clades identified, one of which is the same as in the previous case and the other is a merging of two clades from the previous case.

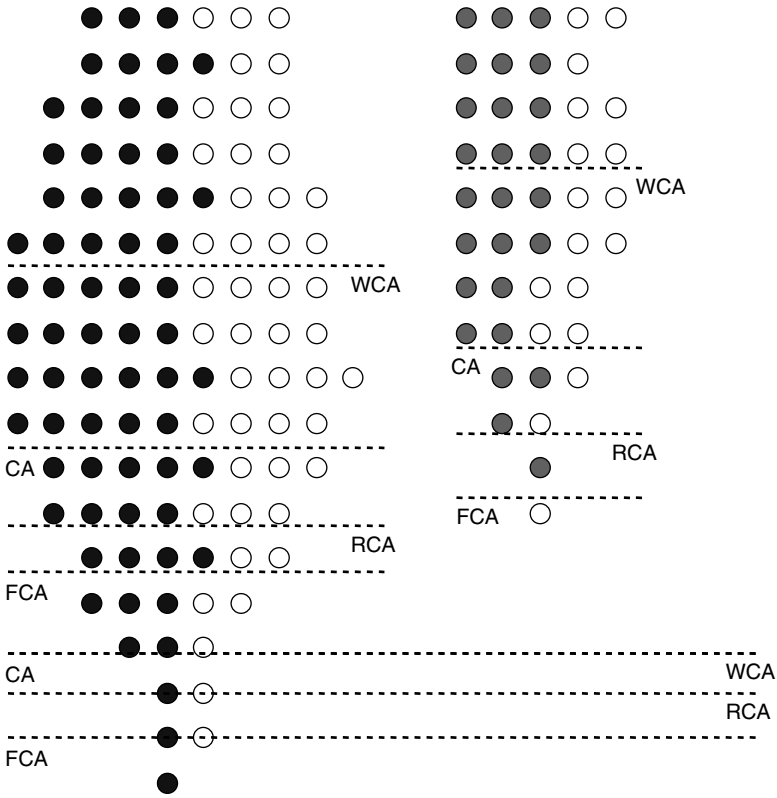


Figure 4.7. Average diversity profile after multiple simulations. The various average most recent common ancestor generations are indicated for each continent and overall. The shaded species indicate the average number of new species, and thus the unshaded species indicate the average number of continuing species.

The remaining two clades from Figure 4.6a no longer satisfy the minimum size requirement and have been discarded. In Figure 4.6c, the minimum size is increased to 50, and the maximum size to 100. These settings led, as one would expect, to two much larger clades. The true fossil relationships, shown in Figure 4.6d, show that there are in fact very deep splits in the lines leading to living species, deeper than indicated by the reconstruction. Therefore, the true situation is best captured by the maximally *splitting* set of clade parameters used to generate Figure 4.6a.

Compared with the 11 true migrations shown in Figure 4.4 and listed on p. 48, the reconstruction identifies 10 migrations: one from *A* to *B*, three from *A* to *C*, one from *A* to *D*, two from *C* to *A*, one from *C* to *B*, one

from *D* to *A* and one from *D* to *B*. The reconstruction thus fails to recover two migrations from *A* to *B*, one from *C* to *B* and one from *D* to *C*, while mistakenly adding two migrations from *C* to *A* and one from *D* to *B*. The average time of migration is at generation 10 for both the reconstruction and the true simulation, as would be expected from the settings, which constrained migrations to occur between generations 5 and 15 with equal probability.

4.3 Advanced features

The more advanced features of the simulation include the more complex features of the mutation model, i.e. reversal penalty and non-linear mutation rate, modelling interbreeding and the consequent merging of lineages, as well as providing the ability to average multiple runs with the application of diversification constraints. The following results provide the simplest possible introduction to these features, which will be central to the much more complex and detailed simulations whose results form the next two chapters.

Starting with average multiple runs, and using the settings employed in the migration example (Section 4.2) but this time with migrations between two continents only, simulations were repeatedly run until 1000 simulations had been obtained with the number of living species at least 4.

The CA was, on average, found at generation 4, and this value was in agreement with the generation determined by the Wagner reconstruction. The FCA was on average at generation 2, but the fossil reconstruction averaged a slightly more recent value of generation 3 (see Figure 4.7).

Important average figures involving the number and accuracy of the connections included 35% of fossils lying on extant lineages, compared with 46% as hypothesised by the fossil reconstruction. The fossil reconstruction was 60% accurate in its connecting fossils to their immediate fossil parent, and this value rose to 72% when considering living species only.

Finally, the clade identification averages showed 52% of identified clades being truly monophyletic, and 47% polyphyletic. Among the polyphyletic cases, 99.3% of true descendant species were found, but 55% of the species included were not actually ancestral to the clade parent. Considering current species only, 90% of those living species included in a clade were truly members of a monophyletic group, contrasting markedly with the results from the Wagner reconstruction where only 31% of current species in a clade were considered members of a monophyletic group.

The addition of non-hereditary characters makes no difference to the diversity profile but does increase the overall character diversity contained therein,

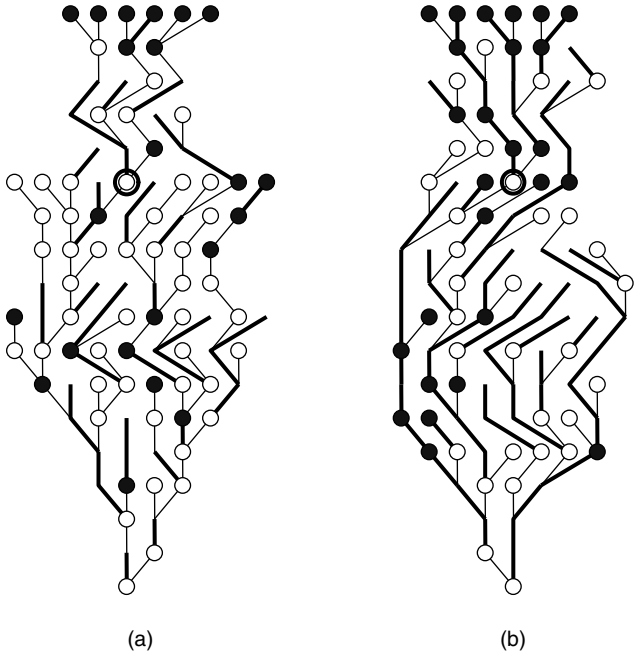


Figure 4.8. (a) A 50% chance of 1 character changing per species per generation; (b) a 10% chance of between 1 and 5 characters changing per species per generation. The most recent common ancestor is circled in both cases.

leading to a reduction of the accuracy of both reconstructions. When 25% of the morphological characters were made non-hereditary, representing 8 different states with a 20% chance of change, the clade identification results were substantially degraded. Less than 14% of clades identified were truly monophyletic in this case, and for current species, only 60% were considered members of a monophyletic group. Increasing the chance of environmental character-state change to 50% drops these figures even further, down to only 2.8% of identified clades being truly monophyletic, and less than 50% of current species in the clades truly belonging to a monophyletic group.

As described in Chapter 3, the mutation model may further include a reversal penalty and a non-linear character mutation rate. Figure 4.8 shows a comparison between two simulations with different mutation models. The mutation model for the simulation shown in Figure 4.8a was a 50% chance of one character changing per species per generation, whereas the simulation shown in Figure 4.8b used a 10% chance of between 1 and 5 characters changing per species per generation.

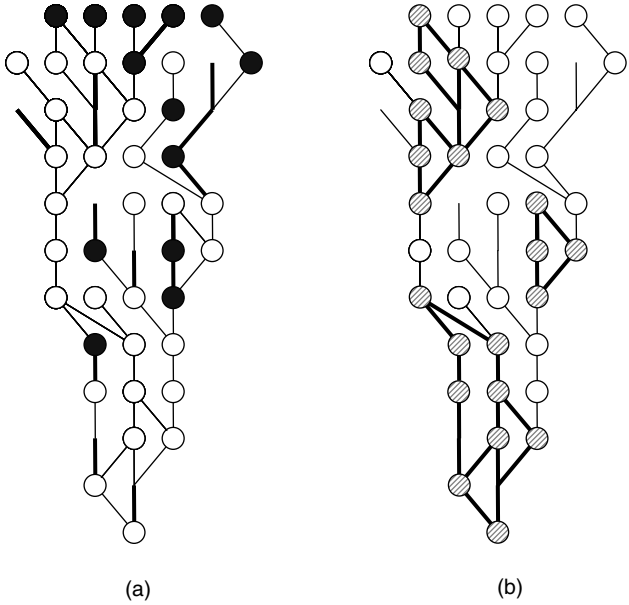


Figure 4.9. An example simulation with interbreeding. Part (b) highlights the splitting and merging of the various lineages, and subspecies are shown shaded.

The most recent common ancestor in both cases occurred five generations back from the present, and the current species character diversity is also essentially the same in both cases, with both simulations having an average current species character pairwise difference of 4.9. However, the very different nature of the species evolution is immediately apparent in the much greater persistence of species in the second case. For example, the initial species persisted for 11 generations and the second species for 12 generations. In contrast, in the first simulation the longest-persisting species managed only four generations. Overall, the average species duration increases by 67%, from 1.61 generations in Figure 4.8a to 2.69 generations in Figure 4.8b. Another aspect of the difference between the two runs is in the number of species generated: 62 species in Figure 4.8a, but only 39 in 4.8b.

Allowing interbreeding changes the biological identification of the evolutionary units of the simulation from species to subspecies or interbreeding groups. The natural analogue of these units implicitly depends to a large degree on the parameter settings. If the chance of interbreeding and the number of character differences required to prevent interbreeding are both high, then the units can be considered as much more closely integrated groups

than species or subspecies. Conversely, with a small chance of interbreeding and only few character differences being sufficient to prevent interbreeding, the simulation units more closely correspond to subspecies.

Figure 4.9 shows an example of a diversity profile with a 50% chance of interbreeding between any two lineages, provided less than 3% of characters differ and the distance to the most recent common ancestor is less than 3 generations. Two effects of interbreeding are immediately obvious: firstly, it is not possible for the reconstruction to attempt to capture lineage-merging, and secondly the fossil connection accuracy (as measured within the simulation) is increased because there are more correct candidates available.

5 *Simulating diversity*

The results presented in this chapter primarily focus on the most fundamental aspects of the simulation: the generation of character diversity, and subsequent phylogenetic reconstruction. As discussed in Chapter 3, the nature of any particular simulation is determined by setting parameters that control the number of characters and the way in which they can change at each generation, the influence of non-hereditary characters, time-dependent fossilisation, and the splitting and optional merging of lineages. By employing different values for these parameters, the simulation can model many different situations.[1] For the purposes of this book, only runs with direct relevance to hominoid evolution have been studied.

Because, as discussed in Sections 2.1 and 2.2, the number of extant hominoid species is far less than appears to have been the case previously, results from simulations employing profiles that feature a recent reduction in diversity are presented first. Following this, results from simulations of subspecies or interbreeding groups are presented, based on profiles that display an increase in recent diversity (Oxnard, 1997; Oxnard and Wessen, 2001).

Of particular interest are those aspects of the simulation that concern the time of the most recent common ancestor, the degree to which fossils enable this to be determined, and the accuracy of the estimates of this provided by both the fossil and Wagner reconstructions. Also important is the accuracy of the fossil reconstruction with respect to the overall fossil connections, and current species to fossil connections in particular. Finally, there is the accuracy of the grouping of taxa in clades to consider in each case. The impact of different fossilisation rates and non-hereditary character behaviour on all these results is studied in detail in Section 5.3.

The time scale implicit in the simulation may be inferred from the expected number of character mutations per generation, the maximum branching per generation, and the interbreeding parameters. For the species simulations presented below, a time scale of roughly one million years per generation appears to be appropriate, whereas for the subspecies simulations a more likely time scale is one of a few hundred thousand years per generation (Oxnard, 1997; Oxnard and Wessen, 2001).

[1] See, for example, the discussion in Oxnard and Wessen (2001).

Using fossils, extant species and clade diversification models similar to those discussed in Section 1.3, Tavaré *et al.* (2002) estimated that no more than 7% of all primate species that have existed are known from fossils. For the simulations in this chapter, a fossilisation (i.e. species preservation) rate of 10% will generally be used. When allowed to vary, a rate that increases with the recency of the species from a low of 5% to a high of 15% will be the usual range. This is consistent with the value of Tavaré *et al.* (2002), seeing that it is to be expected that the species preservation rate for hominids, and the genus *Homo* in particular, will be higher than the primate average.

5.1 Recent reduction in diversity profiles

Three examples of profiles with a recent reduction in diversity have been chosen: vase-shaped, amphora-shaped (Oxnard, 1997) and mass extinction (Sepkoski and Kendrick, 1993). In each case, the final generation is constrained to have between 4 and 6 species, corresponding to the number of extant hominoid species.

The vase profile has an approximately linear increase in diversity, attaining a maximum around the middle generation of the simulation, followed by a linear decrease until the final generation (generation 25). Figure 5.1 shows a single example run with an extinction function that produces, on average, just such a profile.

A number of interesting features are immediately apparent when comparing the three diagrams in Figure 5.1. The most recent common ancestor (CA) of all current species is the species at generation 13, circled in all three figures. It is quite rare that the CA is also the fossil common ancestor (FCA), but being so in this case makes the following discussion of the reconstruction process and its inherent pitfalls all the more informative. Comparing Figure 5.1b and c, it can be seen that the reconstruction identified the true common ancestor fossil as ancestral to only two of the five current species. The other three current species were identified as being descendants of another fossil at generation 14, rather than of the dashed-circled fossil at generation 16, which in turn descends from the CA at generation 13. (This link was correctly identified by the reconstruction.) This reconstruction error deeply obscures the true relationships between the living species, creating a very deep split in the phylogeny, as indicated by the dashed vertical lines. According to the reconstruction, the generation of the fossil common ancestor of all existing species (RCA) is right back at generation 5,

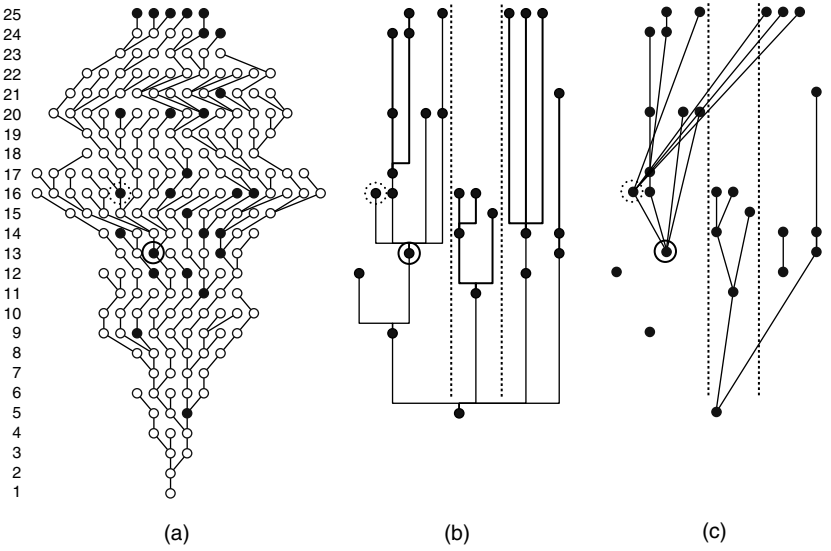


Figure 5.1. (a) A single run with the vase profile; (b) the reconstruction; and (c) the true fossil relationships. The fossil at generation 13, enclosed by a continuous circle, is a true common ancestor of all current species. The fossil at generation 16, enclosed by a dashed circle, is a common ancestor of four of the five current species, but this is not recognised by the reconstruction.

or, after adopting our approximate time scale, *8 million years earlier than the true value*! The Wagner reconstruction calculates the common ancestor (WCA) to be at generation 11, in quite close agreement with the true value.

So how does the reconstruction get the current species connections so wrong? The answer lies in the non-hereditary characters. The character vectors for the three current species in the same clade, all connecting to the wrong ancestor, are

$$S_1: 0100\,1110\,1011\,0100\,0010\,0000\,1000\,0000$$
$$S_2: 0100\,1110\,1111\,0100\,0010\,0010\,1000\,0000$$
$$S_3: 0100\,1110\,1111\,0100\,0010\,0000\,1000\,1000$$

These three species are obviously closely related, as they share many derived characters. The two candidate ancestors are at generation 16 (the true fossil

ancestor) and generation 14 (the reconstructed ancestor), and these have character vectors:

A_{16}: 0010 0011 0011 0100 0010 0000 1000 1000
A_{14}: 0100 1110 0011 0000 0110 0000 1000 0000

The matrix of character differences is thus:

	S_1	S_2	S_3
A_{16}	7 (6)	9 (6)	7 (7)
A_{14}	3 (8)	5 (8)	5 (8)

the value in brackets being the number of shared 1s, and the reconstruction chooses the species at generation 14 as the more likely ancestor since, for all three current species, it has the least differences and the most shared 1s.

However, unknown to the reconstruction, the first eight characters are, in fact, not hereditary. It is clear that the three living species share an environment with the species at generation 14, and not with their ancestor at generation 16. If we exclude these characters, the matrix of differences becomes:

	S_1	S_2	S_3
A_{16}	2 (5)	4 (5)	2 (6)
A_{14}	3 (4)	5 (4)	5 (4)

and the true ancestor at generation 16 is now much more apparent, having the least differences and the most shared 1s in all three cases.

Despite the above significant error, overall the reconstruction is very accurate, with 72% of all true fossil connections being correctly identified.

A final important feature demonstrated by this simulation is the lack of fossils on extant lineages: only 22% of fossils truly lie on extant lineages, increasing to 40% according to the reconstruction. This is despite the high fossilisation rate of a constant 10% being applied.

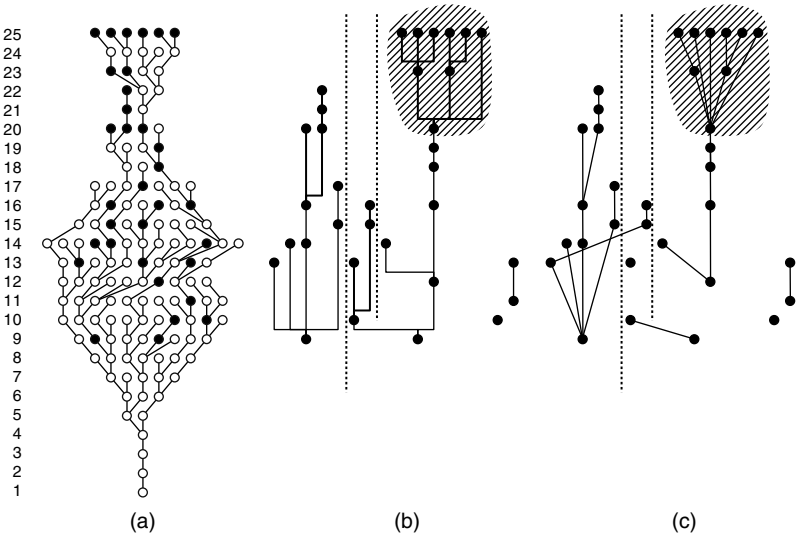


Figure 5.2. (a) A single run with an amphora profile, showing a shallow phylogeny of living species; (b) the reconstruction; and (c) the true fossil relationships. The clade containing all current species is shaded in both fossil figures.

An amphora profile has a long 'neck' of reduced diversity after an initial increase. Figure 5.2 shows one run with an amphora profile, where the ancestor of all living species arose quite late in the overall evolution, well into the neck region of the diversity profile. A very high rate of fossilisation was applied in this simulation, varying from an initial 10% to 30% at generation 25. The result of this is seen in the large number of recent fossils, and the corresponding accuracy in the clade identification. The current species clade is truly monophyletic, with a connection accuracy of 75%: six of the eight connections are correct, but two of the current species are connected to a too recent ancestor. The second clade identified is also monophyletic, with 100% connection accuracy. Only the earliest clade is an artefact of the reconstruction process.

The deep division in the fossil tree, indicated by the dotted vertical lines, is accurately recovered by the reconstruction. All current species are from one branch of this division, so the common ancestor calculations are consistent, accurate, and relatively recent: CA at generation 21, FCA and RCA at generation 20. However, the Wagner common ancestor (WCA) is significantly earlier than these values, seven generations further back in fact. The average pairwise character difference for the current species is only 3.9 characters,

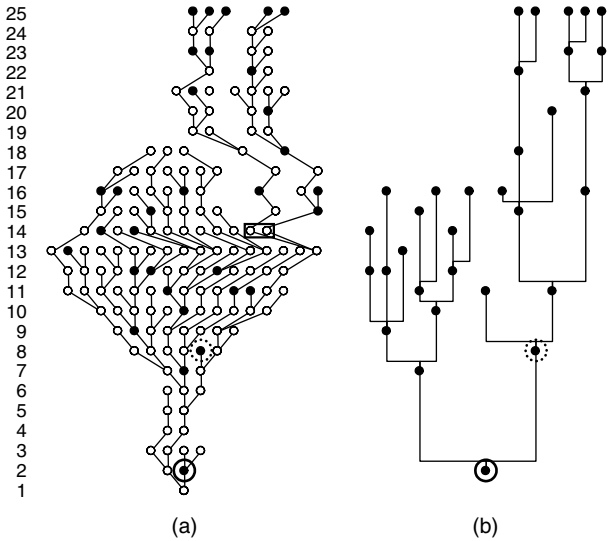


Figure 5.3. A single run with an amphora profile, showing a deep split in the phylogeny of living species; (a) highlights the two evolutionary paths leading to the current species, and (b) shows the fossil reconstruction, in which the deep split is also apparent. The boxed species show the point of divergence of the two lineages leading to current species. The species enclosed by a dashed circle is incorrectly identified as the most recent fossil common ancestor of all current species, and the species enclosed by a continuous circled species is incorrectly identified by the reconstruction as a common ancestor of all fossils.

consistent with a recent common ancestor. Similar to the vase example above, very few fossils lie on extant lineages. Only six of the 27 or 22% of fossil species do so, increasing to eight (or 30%) in the reconstruction.

Figure 5.3 shows a second run using exactly the same parameters as for Figure 5.2. In this case, the simulation resulted in an amphora profile where all the living species are on two quite deep and distinct lineages. The two surviving lineages are shown separated from the remainder of the species in Figure 5.3a, with the two species at the point of divergence highlighted. This division is accurately recovered by the reconstruction (Figure 5.3b), and the true CA at generation 13 is also well recovered by both reconstructions with the RCA at generation 11 and the WCA at generation 12. In this case, the average pairwise character difference for the current species is 8.3 characters, more than twice that of the previous amphora simulation, and, as with the common-ancestor generations, more in keeping with the results of the vase profile simulation shown in Figure 5.1.

The reconstruction finds a common ancestor of all fossils, circled in Figure 5.3, but this identification is in error. The species identified actually has only two descendants, on lineages persisting no further than the next generation. However, it is a sister species to a true common ancestor, and thus the reconstruction, given no other fossil species for several subsequent generations, is unable to discriminate. A similar situation is apparent with the species enclosed by a dashed circle in the figure. This fossil is incorrectly identified as the most recent fossil common ancestor of all current species, when in fact it became extinct immediately.

One side-effect of the high fossil density, seen clearly in this simulation, is that nearly all terminal taxa have been assigned to a clade. However, none of the clades in this case is truly monophyletic. Three of the four are actually polyphyletic groups, although in each case all true ancestors of the parent species are included. The fourth clade is paraphyletic, with 80% of descendant species included. The success rate when considering the current species in each clade alone was much better, with both groupings being truly monophyletic, i.e. with their most recent ancestor ancestral to no other current species.

Another situation that leads to a reduction in diversity is a mass extinction, where after a period of increased diversification there is a sudden and drastic drop. Figure 5.4 shows one run with such a profile. There is a pronounced mass extinction between generations 18 and 19, with the number of species dropping from 52 to 11. This has a marked effect on the diversity of present species because only 1 of 52 species at maximum diversity lies on a lineage that survives to the current generation: the species enclosed by a dashed circle in Figure 5.4a. This means that all other derived characters at the diversity maximum are lost.

The extinction leads to a recent common ancestor for all living species, at generation 20, and a fossil common ancestor at generation 19 (circled in the figure), but both reconstructions perform poorly in this regard. The fossil reconstruction, shown in Figure 5.4b, does not identify any fossil as a common ancestor of all current species, and in fact indicates very deep splits in the phylogeny: three-way splits despite only five current species (indicated by the vertical dashed lines). This is despite the fact that the most recent common ancestor is actually a fossil, and a very recent one at that! The large number of connections crossing the dividing lines in Figure 5.4c show the degree to which these divisions are an artefact of the reconstruction process. The Wagner reconstruction similarly results in a very deep phylogeny, producing an estimate for the common-ancestor generation right back at generation 4.

As for the vase example above (Figure 5.1), understanding the impact of the non-hereditary characters is crucial to understanding the cause of the

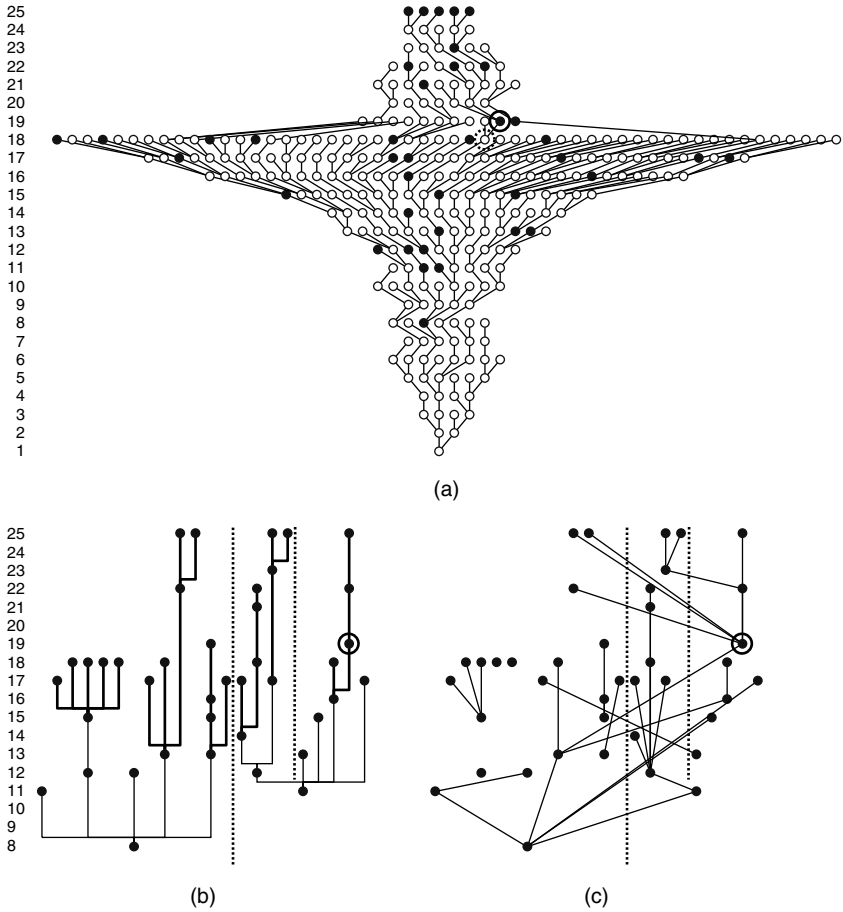


Figure 5.4. Single run with a mass extinction. (a) The full simulation; (b) the fossil reconstruction; (c) the true fossil relationships. Generation numbers are as shown. The dashed circle at generation 18 encloses the only species from the time of maximum diversity whose lineage survives to the current time; the continuous circle at generation 19 encloses the true fossil common ancestor. This is missed by the reconstruction, which instead spuriously identifies the deep divisions indicated by the dashed vertical lines.

errors in reconstruction. In this example, it is the significant environmental differences between recent species, as may arise because of an increased availability of new environments after a mass extinction, that affects the fossil reconstruction.

The character vectors for the five current species are shown below, with the non-hereditary characters shown in brackets.

S_1: (0010 0110) 0000 0000 0001 1110 0100 0000
S_2: (0001 1100) 0001 0000 0001 0101 0100 0000
S_3: (0000 1000) 0001 0000 0001 0101 0100 0000
S_4: (0111 1110) 1001 1001 0001 0000 1000 0000
S_5: (0111 1110) 1001 1001 0001 0000 0000 0000

The character sites at which these species differ are each indicated by • in the following string.

(○●●● ●●●○) ●○○● ●○○● ○○○○ ●●●● ●●○○ ○○○○

In non-hereditary characters, these species differ at 6 of the 8 sites (75%), whereas for hereditary characters they differ at 10 of the 24 sites (42%). Therefore, non-hereditary character differences account for more than one third of the entire difference, leading to the reconstruction difficulties observed.

Although it may seem likely that the non-hereditary characters are also responsible for the problems with the ancestor generation as calculated by the Wagner reconstruction, this is not actually the case. Even if non-hereditary characters are removed when the Wagner reconstruction is applied to the current species, the result is still a WCA occurring at generation 4. The real reason is more fundamental to the Wagner reconstruction process. The hypothetical common ancestor taxon according to the Wagner reconstruction has (hereditary) characters

0000 0000 0001 0000 0000 0000.

This is very close to the outgroup taxon *O*, and thus the time estimate is close to the start of the simulation. The true common ancestor has characters

1001 0001 0001 0000 0100 0000.

Although it is surprisingly primitive for a species 19 generations into the simulation (only 5 of the 24 characters are unambiguously derived; see the discussion in Section 3.1.2) it is obviously far more derived, and thus much later, than the hypothetical Wagner taxon above. So a combination of the loss of diversity at the mass extinction and the rapid evolution thereafter

causes the Wagner reconstruction to be unable to properly determine primitive states, and it is therefore unable to accurately determine the common ancestor generation.

The above examples highlight the important fact that when dealing with random processes any individual run can be quite different from the average. In order to study the average features of these profiles, many simulations, constrained to ensure consistency, must be studied. A constraint of between 4 and 6 current species was applied and 1000 simulations were run with each of these profiles. The results are displayed in Figure 5.5. A fixed fossilisation rate of 10% was used in each case.

Despite a broad similarity in the profiles, there is quite a difference in common ancestry: six generations, or 24%. If each generation corresponds to about a million years, such a big change would, for example, have significant impact on the allocation of fossils to existing lineages. Table 5.1 summarises the runs, and the results presented in Sections 5.3.1 and 5.3.2 examine the dependence of these results on the fossilisation rate employed and on non-hereditary characters, respectively.[2]

The average common-ancestor calculations reveal a number of interesting features. Despite the fact that the actual most recent common-ancestor generation depends a great deal on the particular profile, the true common ancestor is always several generations more recent than suggested by the fossil record. The reason for this is a combination of the sampling effects of fossilisation and the low percentage of connections from fossils to current species. Because the fossil reconstruction only has access to fossil species, its estimate for the common ancestor is similarly affected, but in general is more recent than the true fossil common ancestor. For the vase and amphora profiles the average error is 1.4 generations and 0.5 generations, respectively; for the mass extinction, the generations for the fossil and reconstruction common ancestor differed by only 0.2. A much more detailed study of the dependence of the results on fossilisation rate is presented in Section 5.3.1. Unlike that of a true common ancestor, the existence of a fossil common ancestor is not guaranteed, and the reconstruction also overestimates the frequency of its existence. All of these profiles have few current species, so the likelihood of a common ancestor is relatively high, but in each case the reconstruction found a common ancestor significantly more often than any such ancestor existed.

[2] Although not listed in any of the tables in Chapters 5 and 6, the sample standard deviation s_{N-1} for most values is calculated and included in the text output of the simulation. In general, with $N = 1000$, s_{N-1} was such that the standard error of the means was sufficiently small to make the comparison of average runs statistically unambiguous.

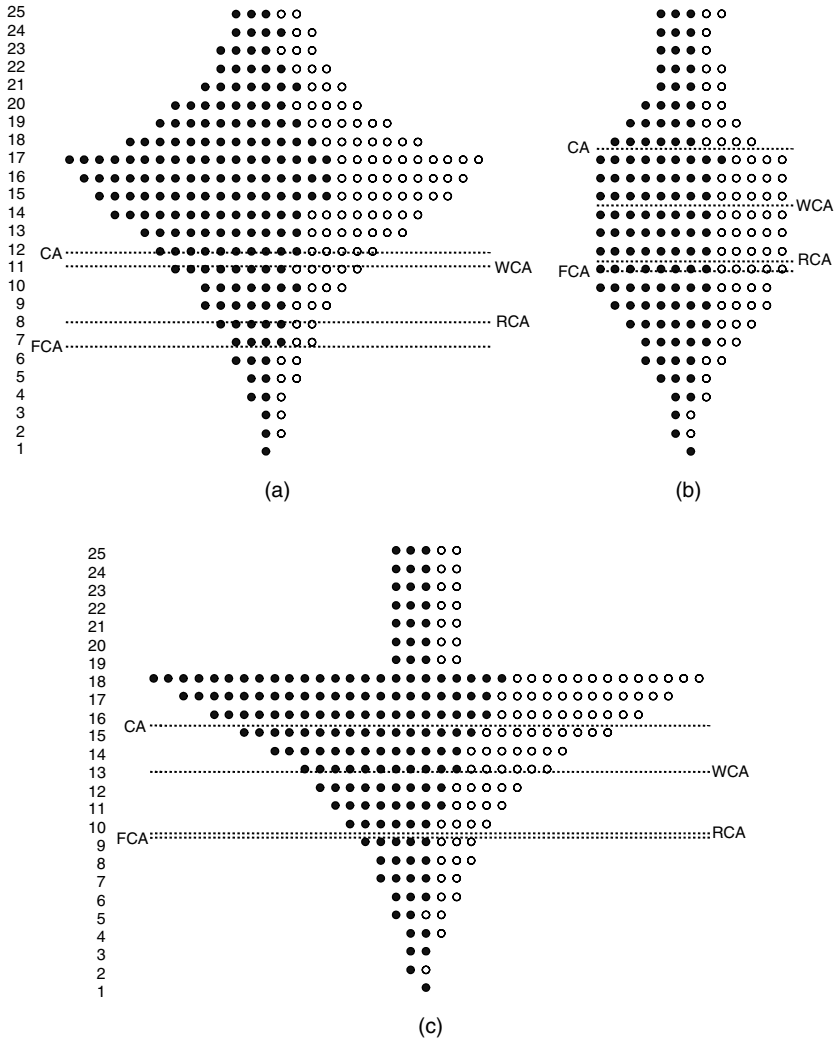


Figure 5.5. Average diversity profiles for the vase profile (a), the amphora profile (b), and mass extinction (c). The generation of the true common ancestor (CA), the Wagner common ancestor (WCA), the fossil common ancestor (FCA) and the reconstruction common ancestor (RCA) are indicated by horizontal lines for each profile, in each case rounded to the nearest generation boundary. The precise values are given in Table 5.1.

Table 5.1. *Average results after 1000 simulations for a vase profile, an amphora profile and a mass extinction*

Current diversity is measured by the average pairwise difference between characters of all current species, and the values in brackets after the fossil common-ancestor generations indicate the percentage of runs where such an ancestor was found. Simulations where there was no fossil common ancestor are not included in the averaging calculation for the relevant common ancestor generation.

Diversity and common ancestry

	Current diversity	True	Wagner	Fossil	Reconstruction
Vase	7.8	12.0	11.2	6.8 (51%)	8.2 (74%)
Amphora	5.9	17.6	14.6	11.0 (71%)	11.5 (83%)
Extinction	6.7	15.4	12.9	9.3 (65%)	9.5 (76%)

Fossils

		Fossils to current		Connection accuracy	
	Fossils	True	Reconstruction	All	Current
Vase	27.7	15%	29%	38%	49%
Amphora	18.0	17%	34%	37%	48%
Extinction	26.6	13%	27%	37%	48%

Clades

	Monophyletic	Paraphyletic (found)	Polyphyletic (found)	Polyphyletic wrong
Vase	16%	8% (68%)	76% (93%)	62%
Amphora	19%	9% (63%)	72% (94%)	63%
Extinction	14%	8% (68%)	78% (93%)	60%

Current species clades

	Reconstruction		Wagner reconstruction	
	Monophyletic	Paraphyletic (found)	Monophyletic	Paraphyletic (found)
Vase	85%	15% (64%)	64%	36% (57%)
Amphora	78%	22% (68%)	62%	38% (61%)
Extinction	77%	23% (66%)	61%	39% (59%)

Notwithstanding the particular difficulties relating to the simulation shown in Figure 5.4, the Wagner reconstruction provides the best estimate of common ancestor generation for all three profiles. The value is in each case far closer to the true value than that obtained from the fossil reconstruction and, by virtue of the reconstruction algorithm, produces a common ancestor estimate every time.

It is quite easy to distinguish these profiles when looking at all species as in Figure 5.5, but the task is much harder when only fossils are available, especially in the case of a single run. For example, try comparing the fossil-only diagrams in Figures 5.1b, 5.2b, 5.3c and 5.4b. However, fossils are all that are available in real life; although the results in Table 5.1 show a similar pattern for all three profiles, the actual values are quite profile-dependent. This serves to illustrate the difficulty in reconstructing a phylogeny from incomplete information, even in such a highly idealised case as provided by the simulation, where the characters are largely unambiguous and fossilisation is high.

The percentage of fossils that lie on existing lineages is also extremely low for all three profiles: less than 20% in all cases. This means that in the fossil reconstruction fewer than 1 in 5 fossils should be assigned to an existing lineage. However, the reconstruction algorithm approximately doubles this value consistently. Extra connections arise in two general ways. The first is via the incorrect interpretation of character similarity as indicating direct ancestry when it is actually the result of either the close relationship of a fossil to a true direct ancestor or of non-hereditary, adaptive similarities (as discussed regarding Figure 5.1 above). The second cause of spurious connections is that non-hereditary, adaptive changes can lead to an overestimate of the mutation rate, and thus too great a tolerance to character difference. The net effect of these errors is usually a more recent RCA than FCA.

The connection accuracy, i.e. the percentage of times a current species or fossil is correctly connected to its immediate fossil parent, is consistent across all profiles, and less than 40% in all cases. The degree to which this is a result of non-hereditary characters or a result of the inherent difficulty in the reconstruction process is discussed in Section 5.3.2. Considering current species alone, the accuracy improves only slightly, to nearly 50%, again independent of profile, despite these profiles having only a few current species.

There is not a great deal of variation in the average results with respect to clade identification. The majority of clades identified are in fact polyphyletic for all profiles; the amphora profile is best for monophyletic clades, but still not particularly good ($<20\%$). The polyphyletic clades do include nearly all the true descendants ($>90\%$ for all profiles) but are significantly overly inclusive, with approximately 60% of included species not actually descended from the parent. This is consistent with the overconnectivity inherent in the reconstruction algorithm.

Considering current species only, the clade identification performs much better, with all profiles scoring greater than 75%. The fossil and Wagner reconstructions for current species show a reasonable degree of agreement, matching in between 61% and 64% of cases.

5.2 Recent maximum of diversity profiles

As discussed in Chapter 3, when interbreeding is added to the model the individual simulation units are more appropriately considered subspecies or interbreeding groups, and so the relevant profiles differ from those of the previous section and the approximate time scale associated with each generation is shorter. In the discussion that follows, the simulation units will be referred to simply as *taxa*. The simulations presented in this section all involve significant merging of lineages, and the particular profiles employed are bowl-shaped (Oxnard, 1997; Oxnard and Wessen, 2001) and logistic (Sepkoski and Kendrick, 1993). The merging probability in these simulations is set to 50%, given a difference of no more than three characters and a common ancestor of no more than three generations back. For the average runs, a linearly varying fossilisation rate of 5% to 15% was employed.

The bowl profile is generated by applying a constant extinction rate of 25%, so the diversity increases with each generation. Figure 5.6 shows a sample run with such a profile. The general increase in diversity with generation is obvious, although somewhat slower than the average for such a low extinction rate. Two very ancient lineages persist to the current generation, having diverged right back at generation 4, highlighted by the shaded area in the figure. Although the divergence of these lineages is ancient, the set of current species belonging to each has a much more recent common ancestor, circled in the figure, 8 and 9 generations back, respectively.

The fossil reconstruction and true fossil connections are shown in Figure 5.7. The reconstruction (Figure 5.7a) accurately recovers the ancient divergence, as indicated by the two shaded regions and the lack of interconnections between them in Figure 5.7b. However, the connection accuracy is only 50%, because the reconstruction tends to connect in error to more recent fossils. The circled ancestor on the right in the reconstruction and its reconstructed descendants illustrate this problem particularly well. There are actually no fossil descendants of this taxon, with nearly all taxa on the right-hand side connecting directly to the earliest fossil in the simulation. However, the reconstruction builds a tree of many levels based on this taxon and its supposed descendants.

The circled ancestor on the left, and the darker shaded segment, illustrate the difficulty in reconstruction when interbreeding occurs. There are two fossil ancestors close in time, but only one can be chosen by the reconstruction algorithm, and this can obscure fossil relationships.

Figure 5.8 shows the results for both a bowl profile and a logistic profile, i.e. a profile in which the diversity increases until an equilibrium value, averaged over 1000 runs. The most immediately obvious feature of these results is how much earlier the common ancestors are occurring in the evolution

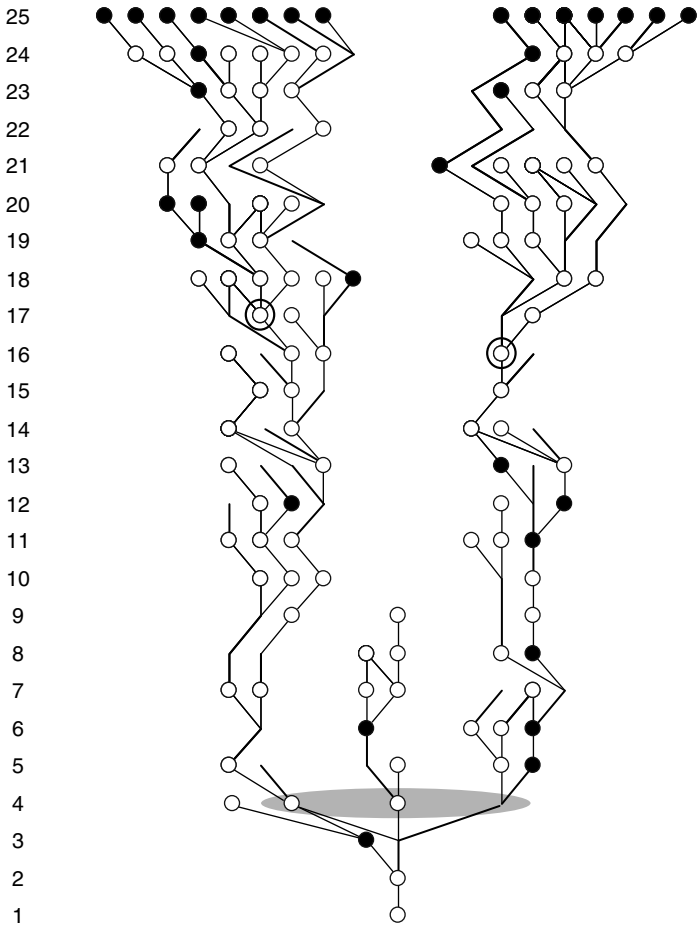


Figure 5.6. A sample single run with a bowl profile, showing the persistence of two ancient lineages. The shading at generation 4 highlights the point of divergence of these lineages; circled in each surviving lineage, at generations 17 and 16 respectively, is the most recent common ancestor of all current species belonging to that particular lineage.

when compared with Figure 5.5, although now there is interbreeding, each simulation step is associated with a smaller time period. Another obvious and interesting difference is the poor performance of the Wagner reconstruction. In the recent minimum simulations, the Wagner estimate consistently gave the best estimate for the common ancestor generation, but in this case it is compromised by the extra diversity of the current species. For these profiles,

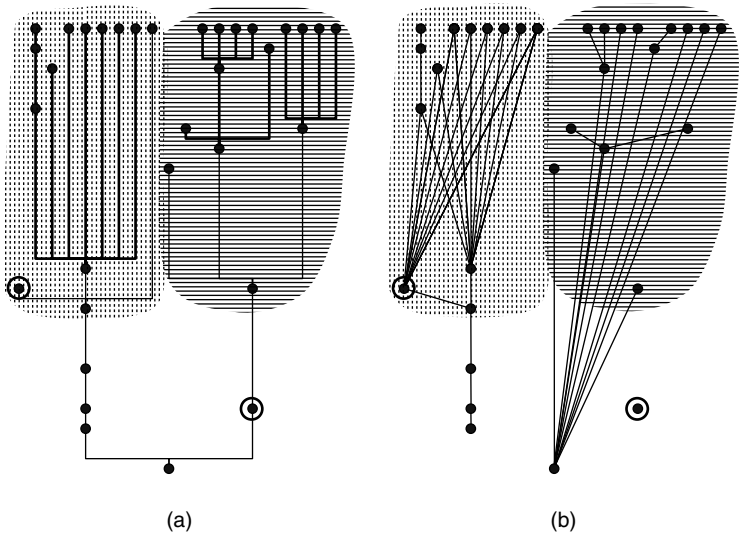


Figure 5.7. Fossils only for the bowl profile run. (a) The fossil reconstruction; (b) the true fossil connections, including merging. The two shaded regions highlight the success of the reconstruction in recovering the two distinct lineages leading to current species. Nevertheless, errors in the reconstruction are apparent, especially as indicated by the missed connections associated with the circled fossils.

the average pairwise character difference is now around 11 characters rather than the 6–8 character difference found in the recent minimum case.

Table 5.2 summarises the runs. For these profiles, the true CA is more recent than any of the other estimates, yet all estimates are very early in the evolution. As was the case for the single-run bowl-profile example discussed above, the existence of a true fossil ancestor is quite rare. For both profiles it occurs in less than one third of simulations, and although the reconstruction substantially overestimates the frequency it still finds one in less than 60% of runs (compared to the 70 + % and 80 + % values for the recent minimum average results). As mentioned above, the Wagner estimate is earliest of all: right back at the start of the evolution (or even earlier)!

There is a much higher percentage of fossils connecting to current species than in the recent minimum case, but still less than 50% despite the high number of current taxa, the effects of interbreeding, and the diversity-increasing nature of the profiles. The reconstruction again overestimates this percentage, by as much as 66%, and the connection accuracy remains poor.

The clade identification performance is even worse than the low figures for the recent minimum profiles, with only 11% and 13% of clades monophyletic.

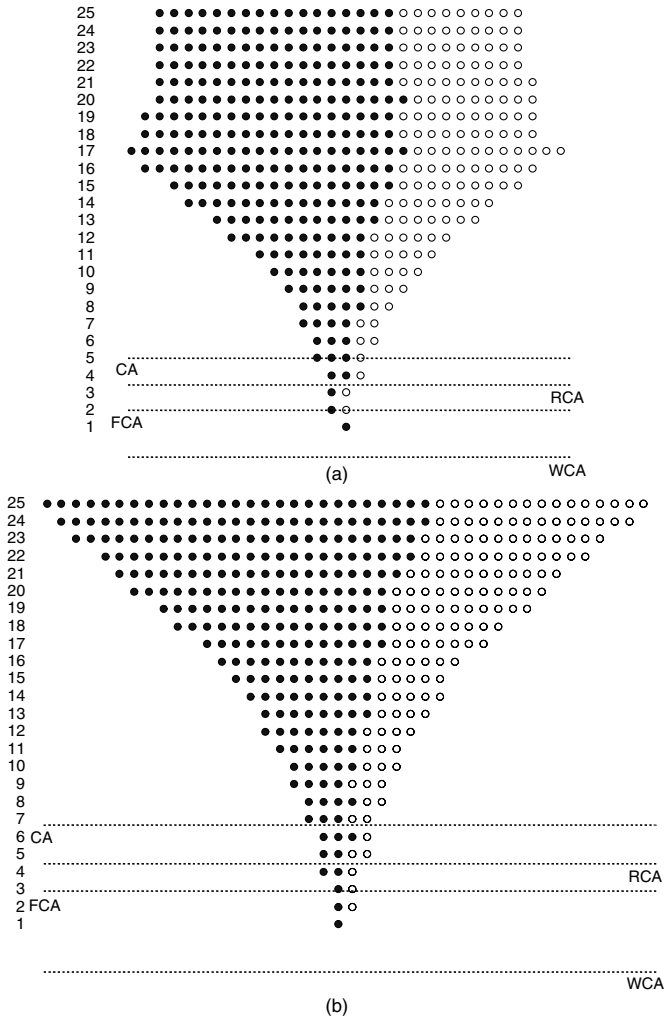


Figure 5.8. Average profiles over 1000 runs for bowl and logistic profiles with interbreeding. The generation of the true common ancestor (CA), the Wagner common ancestor (WCA), the fossil common ancestor (FCA) and the reconstruction common ancestor (RCA) are indicated by horizontal lines for each profile.

The profile for polyphyletic clades is similar to the recent minimum case, with nearly all true descendants found but many additional taxa included. Given the many current taxa in these profiles, the current species clade identification is more problematic, and the performance suffers relative to last time quite

Table 5.2. *Average results after 1000 simulations for a bowl profile and a logistic profile, with 50% chance of closely related lineages merging*

See the headnote to Table 5.1 for more details of the entries.

Diversity and common ancestry

Profile	Current diversity	True	Wagner	Fossil	Reconstruction
Bowl	10.8	6.7	−1.7	3.0 (34%)	4.5 (60%)
Logistic	11.1	5.0	−0.7	2.0 (24%)	3.5 (55%)

Fossils

		Fossils to current		Connection accuracy	
	Fossils	True	Reconstruction	All	Current
Bowl	40.6	49%	72%	42%	45%
Logistic	42.8	35%	58%	41%	46%

Clades

	Monophyletic	Paraphyletic (found)	Polyphyletic (found)	Polyphyletic wrong
Bowl	11%	9% (63%)	81% (88%)	58%
Logistic	13%	10% (65%)	77% (90%)	60%

Current species clades

	Reconstruction		Wagner reconstruction	
	Monophyletic	Paraphyletic (found)	Monophyletic	Paraphyletic (found)
Bowl	38%	62% (37%)	12%	88% (13%)
Logistic	54%	46% (44%)	23%	77% (19%)

substantially. For the fossil reconstruction, the monophyletic percentage is below 40% for the bowl profile and only just over 50% for the logistic profile. The matching between the reconstruction methods for current species clades is not so good for these profiles, although it is hard to interpret the worst Wagner value of 12% for the bowl profile because the fossil reconstruction value is also quite poor in this case. However, even for the logistic profile, the better-performing fossil case, the Wagner performance is exceedingly poor, matching for only 12% of clades, and when there is paraphyly there is agreement on fewer than one in five clade members.

5.3 Studying parameter sensitivity

The simulation results in the previous two sections give an indication of the importance of profile shape when determining common ancestry and general phylogenetic relationships between current species and fossils. However, the important features of fossilisation rate and non-hereditary character change were kept the same in all cases. The sensitivity of the evolution and reconstruction to these parameters is studied in this section, focussing on the most relevant diversity profiles: vase, amphora and logistic. Interbreeding is included in the logistic profile simulations.

Before comparing all these results, it is important to know some baseline figures regarding the number of fossils on extant lineages and the limiting accuracy of the reconstruction algorithms employed. These are provided by running the simulation with a fossilisation rate of 100% (so all species are preserved as fossils) and all characters hereditary. The results are shown in Table 5.3.

The generation estimates for common ancestors are very good, especially for the recent minimum profiles, where they are never wrong by more than a single generation. Because fossils are not involved in the Wagner estimate, a 100% fossilisation rate has no effect on its accuracy, and the accuracy seen here is due to the absence of non-hereditary characters. It might have been expected that the number of additional fossils present would cause the fossil reconstruction algorithm to be too likely to connect, leading to an underestimate of the distance to the common ancestor, but this is not the case. The fossil reconstruction performs well, and when in error for the logistic profile it is actually too far back in time. Of course, the true fossil common ancestor is exactly the true common ancestor in all cases.

A connection accuracy of 95% or more for all three profiles is an excellent result, and shows that fundamentally the reconstruction algorithm is very accurate. The clade identification benefits from the very high connection accuracy, but does not perform quite as well overall. Nevertheless, the results are still very acceptable: for both recent minimum profiles, monophyly is well over 80% and the current species monophyly is approximately 98%, whereas for the recent maximum profile there is monophyly in almost two thirds of cases and current species monophyly in more than 90%. There is very little polyphyly, even for the vase profile where monophyly is less. Interestingly, the more formal Wagner algorithm performs poorly on the clade identification measure, even though here it is being compared with almost perfectly accurate fossil reconstructions. This is especially clear when using the logistic profile, where the Wagner degradation far exceeds the fossil reconstructions difficulties.

Table 5.3. *Summary of simulation results when all characters are hereditary and the fossilisation rate is 100%*

Common ancestor generation				
	True	Wagner	Fossil	Reconstruction
Vase	11.8	11.8	11.8 (100%)	11.7 (100%)
Amphora	17.0	16.0	17.0 (100%)	16.8 (100%)
Logistic	5.4	1.4	5.4 (100%)	4.6 (100%)

Fossils				
	Fossils to current		Connection accuracy	
	True	Reconstruction	All	Current
Vase	15%	15%	95%	97%
Amphora	17%	17%	95%	96%
Logistic	35%	32%	95%	96%

Clades				
	All		Current monophyletic	
	Monophyletic	Polyphyletic	Fossil recon.	Wagner
Vase	82.9%	9.1%	98.4%	64.7%
Amphora	83.5%	8.5%	97.4%	67.5%
Logistic	63.0%	8.6%	90.9%	28.3%

The percentage of fossils leading to current species remains low, especially in the recent minimum cases. This is an important indication that the low value found in Section 5.1 is not due to a scanty fossil record but is a fundamental feature of the evolution according to this model.

5.3.1 Sensitivity to fossilisation rate

To study the sensitivity of the results to the chosen fossilisation rate, 1000 simulations were averaged, based on vase, amphora and logistic profiles for each of four different fossilisation-rate settings: constant 30%, constant 10%, constant 3%, and linearly varying from 5% at generation 1 to 15% at generation 25. The results for the amphora profile are shown in Table 5.4 (the results for the vase profile essentially followed the same patterns, and so

Table 5.4. *Average results for an amphora profile with different fossilisation rates*

Diversity and common ancestry

Fossilisation rate	Current diversity	True	Wagner	Fossil	Reconstruction
30%	6.0	17.3	14.3	15.2 (92%)	10.9 (86%)
10%	5.9	17.5	14.5	10.8 (81%)	10.8 (85%)
3%	6.0	17.2	14.4	5.2 (40%)	9.7 (79%)
5%–15%	5.9	17.6	14.6	11.0 (71%)	11.5 (83%)

Fossils

		Fossils to current		Connection accuracy	
	Fossils	True	Reconstruction	All	Current
30%	57.9	17%	21%	59%	59%
10%	19.4	17%	33%	40%	46%
3%	5.8	17%	54%	22%	29%
5%–15%	18.0	17%	34%	37%	48%

Clades

	Monophyletic	Paraphyletic (found)	Polyphyletic (found)	Polyphyletic wrong
30%	21%	19% (59%)	60% (88%)	58%
10%	17%	10% (64%)	74% (94%)	62%
3%	13%	6% (65%)	82% (98%)	66%
5%–15%	19%	9% (63%)	72% (94%)	63%

Current species clades

	Reconstruction		Wagner reconstruction	
	Monophyletic	Paraphyletic (found)	Monophyletic	Paraphyletic (found)
30%	82%	18% (68%)	69%	32% (60%)
10%	80%	21% (67%)	66%	34% (60%)
3%	73%	27% (65%)	62%	38% (62%)
5%–15%	78%	22% (68%)	63%	37% (61%)

are not presented here). Eight of the 32 characters used were non-hereditary, with a 20% chance of change among eight different states.

For the recent minimum profiles, the generation of the true common ancestor and the Wagner estimate were both independent of the fossilisation rate because fossils are not involved in either case. As expected, the time of the

fossil common ancestor was strongly dependent on the fossilisation rate: for the high fossilisation rate of 30%, the fossil common ancestor only lagged the true common ancestor by a couple of generations, but slipped back several generations as the fossilisation rate decreased. The most surprising result was that the time of the reconstruction common ancestor was largely independent of the fossilisation rate! As the fossilisation rate was reduced from 30% to 3%, the generation of the fossil common ancestor changed from generation 15 to generation 5, but the reconstruction common-ancestor estimate ranged only between 9.7 and 11.5 over all four cases. This certainly indicates a flaw in the reconstruction algorithm, but quite a realistic one. The reconstruction can only proceed on the basis of possible relatedness, and will always err on the side of connection if a possible ancestor fossil is available. These results nicely illustrate the two directions of error in the reconstruction. In the high-fossilisation case, the presence of non-hereditary characters cause the reconstruction to be unable to take full advantage of the extra information provided by the high degree of fossilisation. When there are many fossils to choose between, the relative importance of matching non-hereditary characters increases and leads to connection errors. Because the reconstruction algorithm places more emphasis on character-matching than on closeness in time, the estimate tends to be pushed back in time. In the low-fossilisation case there is generally less close matching of fossil characters, and so the rate of mutation tends to be overestimated and the connections err in the direction of connecting too recent fossils.

The percentage of fossils leading to current species in these results is constant and, as indicated by the 100% fossilisation rate results in Table 5.3 above, agrees with the true value of 17%. However, the reconstruction fossils to current species percentage is highly dependent on fossil availability. For the 3% fossilisation rate, this percentage increases to more than three times the true value, with more than half of all fossils placed on extant lineages. As with the common ancestor estimation errors discussed above, this indicates a flaw in the algorithm, but again quite a realistic one. When sufficient fossils are available, the reconstruction algorithm connects very accurately, but the connection accuracy drops to only just over 20% for all fossils in the 3% fossilisation case, and remains below 30% even when only current species are being considered. This is because, when there are only relatively few fossils available, the algorithm is unable to accurately resolve relatedness and errs in the direction of overconnection.

The accuracy of the clade identification is also reduced, owing to the drop in connection accuracy described above, and well over half of all clades identified are polyphyletic for both the vase and amphora profiles. Across all measures, the lower the fossilisation rate, the poorer the performance.

For current species clades the trend is similar but less pronounced, and in over 70% of cases the current species grouped together in a clade are in fact all the extant members of a monophyletic group for both profiles and all fossilisation rates. As in most of the simulations presented so far, the Wagner current species results are worse, but nevertheless match the fossil reconstruction results in over 60% of cases.

The results for the logistic profile with interbreeding are summarised in Table 5.5. The logistic profile results display a general dependence on the fossilisation rate similar to that of the amphora and vase results, although the particular features described in Section 5.2 remain apparent. Because the common-ancestors (true and according to both reconstructions) are much earlier in the simulation, there is less scope for variation, but the impact of the changing fossilisation rate is nevertheless clearly seen by looking at the drop in the percentage of runs in which a fossil common ancestor was found, from 70% for 30% fossilisation, down to only 15% for 3% fossilisation. For the recent minimum profiles, the percentage of runs in which a reconstruction common ancestor was found varied very little with fossilisation rate although the percentage for the fossil common ancestor varied widely. It is interesting to see that this behaviour remains in the recent maximum case, even when the true fossil common ancestor found percentage drops so low.

The number of fossils connected to current species is again essentially constant and close to the true value for all fossilisation rates employed, but for the low fossilisation rate in particular the reconstruction connects far too many fossils to extant lineages: 80%! This is a reflection of the fact that the connection accuracy is a very low 24% in this case. A further consequence is that all the measures of clade monophyly also show very poor performance with the low fossilisation rate.

Because the Wagner reconstruction uses only current taxa, it is not affected by the changes in fossilisation rate. The percentage values for current clade matching here simply follow the fossil reconstruction values: as the fossil reconstruction degrades, the unaffected Wagner reconstruction matches less and less.

Availability of fossils is obviously very important for accurate reconstruction, but there is no point in running simulations where so many fossils are present that the results bear no relation to reality. Because the estimate of the reconstruction common-ancestor generation is quite good for the 5%–15% fossilisation function, and, although quite high, this fossilisation rate is not so high as to remove from the simulations the very kinds of problem they are attempting to model (Tavaré *et al.*, 2002), it was adopted as standard for all following results.

Table 5.5. *Average results for a logistic profile with different fossilisation rates*

Diversity and common ancestry

Fossilisation rate	Current diversity	True	Wagner	Fossil	Reconstruction
30%	11.1	5.0	−0.5	3.7 (70%)	3.0 (67%)
10%	11.1	4.8	−0.8	2.2 (38%)	3.5 (63%)
3%	11.1	5.0	−0.9	1.5 (15%)	4.1 (54%)
5%–15%	11.1	5.0	−0.7	2.0 (24%)	3.5 (55%)

Fossils

		Fossils to current		Connection accuracy	
	Fossils	True	Reconstruction	All	Current
30%	118.0	34%	42%	60%	60%
10%	40.5	34%	58%	42%	42%
3%	11.8	35%	80%	24%	26%
5%–15%	42.8	35%	58%	41%	46%

Clades

	Monophyletic	Paraphyletic (found)	Polyphyletic (found)	Polyphyletic wrong
30%	18%	20% (60%)	62% (84%)	58%
10%	12%	10% (62%)	78% (90%)	60%
3%	6%	3% (65%)	91% (95%)	65%
5%–15%	14%	10% (65%)	77% (90%)	60%

Current species clades

	Reconstruction		Wagner reconstruction	
	Monophyletic	Paraphyletic (found)	Monophyletic	Paraphyletic (found)
30%	69%	31% (50%)	32%	68% (19%)
10%	53%	47% (43%)	22%	78% (19%)
3%	30%	70% (35%)	9%	91% (20%)
5%–15%	54%	46% (44%)	23%	77% (19%)

5.3.2 Sensitivity to non-hereditary characters

In this section, the sensitivity of the results to the presence of, and changes in, non-hereditary characters is studied in a manner similar to the study of sensitivity to fossilisation rate above. Again 1000 simulations were run and

averaged for the three main profiles, vase, amphora and logistic, and for each profile four different kinds of non-hereditary character behaviour were simulated. Recall that, unlike hereditary characters, the non-hereditary characters are not independent of each other, but rather are sets of non-hereditary adaptive characters, each representing a particular environment–lifestyle combination. The simulation then models a change in this environment, with consequent change in the non-hereditary characters, rather than modelling change in the non-hereditary characters directly. The three parameters involved are the number of non-hereditary characters, the number of distinct adaptive states these characters represent, and the probability of change. The results presented in this section include the case where all characters are hereditary, providing a baseline result for comparison, and three different settings of these parameters:

1. 8 of 32 characters non-hereditary, representing 8 different adaptive states with a 20% probability of change;
2. 16 of 32 characters non-hereditary, representing 16 different adaptive states with a 20% probability of change; and
3. 8 of 32 characters non-hereditary, representing 8 different adaptive states with a 50% probability of change.

The results for the amphora profile are shown in Table 5.6; the vase profile results showed a similar pattern.

Contrary to the fossilisation case presented in the previous section, the Wagner common-ancestor generation is significantly influenced by the nature of the non-hereditary characters in the simulation, as is the reconstruction estimate. However, this time the true fossil common-ancestor generation is unaffected. So, in combination, the fossilisation rate and non-hereditary character parameters exert significant influence on all three common ancestor estimates.

The number of fossils connected to current species is also independent of the non-hereditary character modelling, as indeed are most of the direct fossil measures, because non-hereditary characters have no influence on the probability of fossilisation. Although the total number of fossil connections increases, the reconstruction percentage of fossils connected to current species changes only slightly as the importance of non-hereditary characters increases, and both connection accuracy percentages drop: down to as low as 31% after being near 50% when all characters are hereditary. Monophyly among the identified clades thus also suffers substantially, although the current species clades continue to do quite well.

Using the logistic profile and introducing interbreeding gives the average results summarised in Table 5.7. Many features of these results echo the above examples, but are often more pronounced. There is a similar degradation in the

Table 5.6. *Average results for an amphora profile with different non-hereditary character modelling*

A label of *m*/*n*/*p*% represents a run using *m* non-hereditary characters, representing *n* distinct adaptive states, with probability of change *p*%.

Diversity and common ancestry

	Current diversity	True	Wagner	Fossil	Reconstruction
All hereditary	4.3	17.5	16.3	10.8 (72%)	13.6 (88%)
8/8/20%	5.9	17.6	14.6	11.0 (71%)	11.5 (83%)
16/16/20%	8.2	17.8	12.7	11.0 (70%)	7.9 (71%)
8/8/50%	6.4	17.3	13.5	10.7 (69%)	11.1 (88%)

Fossils

		Fossils to current		Connection accuracy	
	Fossils	True	Reconstruction	All	Current
All hereditary	18.1	17%	29%	46%	62%
8/8/20%	18.0	17%	34%	36%	48%
16/16/20%	17.7	17%	38%	31%	40%
8/8/50%	18.5	17%	38%	31%	43%

Clades

	Monophyletic	Paraphyletic (found)	Polyphyletic (found)	Polyphyletic wrong
All hereditary	36%	4% (69%)	61% (99%)	59%
8/8/20%	19%	9% (63%)	72% (94%)	63%
16/16/20%	7%	12% (63%)	81% (92%)	63%
8/8/50%	12%	11% (64%)	78% (92%)	63%

Current species clades

	Reconstruction		Wagner reconstruction	
	Monophyletic	Paraphyletic (found)	Monophyletic	Paraphyletic (found)
All hereditary	89%	11% (69%)	71%	29% (61%)
8/8/20%	78%	22% (68%)	63%	37% (61%)
16/16/20%	73%	27% (64%)	65%	35% (58%)
8/8/50%	70%	30% (64%)	59%	41% (61%)

Wagner common ancestor estimation (already poor for this profile). There is some indication of an effect on the reconstruction common ancestor estimate as well, but the results are not conclusive, possibly as a result of the generation already being so early in the evolution. Nevertheless, as the impact of the

Table 5.7. *Average results for a logistic profile with different non-hereditary character modelling.*

A label of *m*/*n*/*p*% represents a run using *m* non-hereditary characters, representing *n* distinct adaptive states, with probability of change *p*%.

Diversity and common ancestry

	Current diversity	True	Wagner	Fossil	Reconstruction
All hereditary	9.9	5.2	1.1	2.1 (27%)	2.9 (43%)
8/8/20%	11.1	5.0	−0.7	2.0 (24%)	3.5 (55%)
16/16/20%	13.1	4.8	−2.7	1.9 (23%)	3.1 (52%)
8/8/50%	11.1	4.6	−1.7	1.9 (23%)	4.8 (76%)

Fossils

		Fossils to current		Connection accuracy	
	Fossils	True	Reconstruction	All	Current
All hereditary	43.9	36%	51%	53%	61%
8/8/20%	42.8	35%	58%	41%	46%
16/16/20%	44.1	34%	62%	33%	35%
8/8/50%	44.9	34%	61%	33%	38%

Clades

	Monophyletic	Paraphyletic (found)	Polyphyletic (found)	Polyphyletic wrong
All hereditary	33%	7% (65%)	60% (96%)	58%
8/8/20%	14%	10% (65%)	77% (90%)	60%
16/16/20%	4%	8% (59%)	89% (84%)	61%
8/8/50%	7%	9% (60%)	84% (87%)	61%

Current species clades

	Reconstruction		Wagner reconstruction	
	Monophyletic	Paraphyletic (found)	Monophyletic	Paraphyletic (found)
All hereditary	75%	25% (57%)	19%	81% (21%)
8/8/20%	54%	46% (44%)	23%	77% (19%)
16/16/20%	43%	57% (29%)	26%	74% (17%)
8/8/50%	44%	56% (41%)	23%	77% (18%)

non-hereditary characters increases, the reconstruction is forced to be more tolerant in making its connections and the reconstruction common-ancestor generation becomes correspondingly more likely to be found, and found more recently.

The percentage of fossils on extant lineages is again independent of the nature of the non-hereditary characters, and the reconstructed value only slightly dependent. In the absence of non-hereditary characters, the connection accuracy is very similar for all three profiles. However, for the logistic profile the impact of non-hereditary characters is more severe than is the case with the vase and amphora profiles, dropping below 30% in the worst case.

Clade identification performance remains very poor for this profile. For both non-hereditary characters and fossilisation rate, this is the most sensitive area, yet it is a crucial one for phylogenetic reconstruction. For the runs where 16 of the 32 characters are non-hereditary, the percentage of identified clades that were truly monophyletic was only 4%, and 90% of clades were polyphyletic. With all characters hereditary, the performance was significantly better, with 33% of identified clades being truly monophyletic, i.e. better than half the rate when fossilisation is 100% (see Table 5.3).

The impact of the non-hereditary characters is even greater on the current species clade results. There is a much greater reduction in monophyly, with a nearly 50% drop from the best to the worst case, as opposed to the slight drop found for the vase and amphora profiles. The results in Table 5.3 show a best possible monophyly rate of 91% for current species clades, but this drops to between 43% and 54% for all three cases with non-hereditary characters presented in Table 5.7.

The Wagner current species monophyly results here show a marked difference from earlier patterns, with the lowest degree of correspondence with the fossil reconstruction clades for the case where all characters are hereditary. As the non-hereditary characters are increased in number and variation, the degree of correspondence between the two reconstructions increases even though the accuracy of each is actually reduced, showing that they are both being affected by the non-hereditary characters in similar ways.

6 *Simulating migration*

With the results from the previous chapter as a basis, additional factors can be added to the simulation, allowing it to be applied to the study of problems directly relevant to human origins. The two principal additional factors employed in this chapter are migration and selective advantage.

Of the many different and interesting ways to model migration and selective advantage, the simulations presented here focus on the following situations:

1. with an amphora profile, study the unrestricted migration of species between two continents (initial population on one only);
2. with an amphora profile, study migration between three continents, where the migration patterns approximate those of the hominoids, outlined in Section 2.1;
3. with a logistic profile with interbreeding, study migration between two continents (initial population on one only) with migrations restricted in time to four different six-generation windows;
4. with the same profile as above, study 'spatially' restricted migration where one continent acts purely as a source continent and the other two purely as sinks (i.e. destinations only; no migration from these continents) and see the effect of different degrees of selective advantage between the source and sink continents;
5. still with the same profile as above, study unrestricted migration between all four continents, each with initial population but with species originating on one particular continent having a significant selective advantage over all the others.

6.1 Species migration with an amphora profile

The simulations presented in this section are designed to allow study of the most simple case of species migration, with an amphora profile and only two continents. A high and varying fossilisation rate of 10%–30% is used because, although relatively few species are produced in such a simulation, the special interest in human species in particular leads to a greater number of findings, and the variation covers the fact that more recent fossils are more likely to be found. Following the results in Section 5.3.2, the default non-hereditary

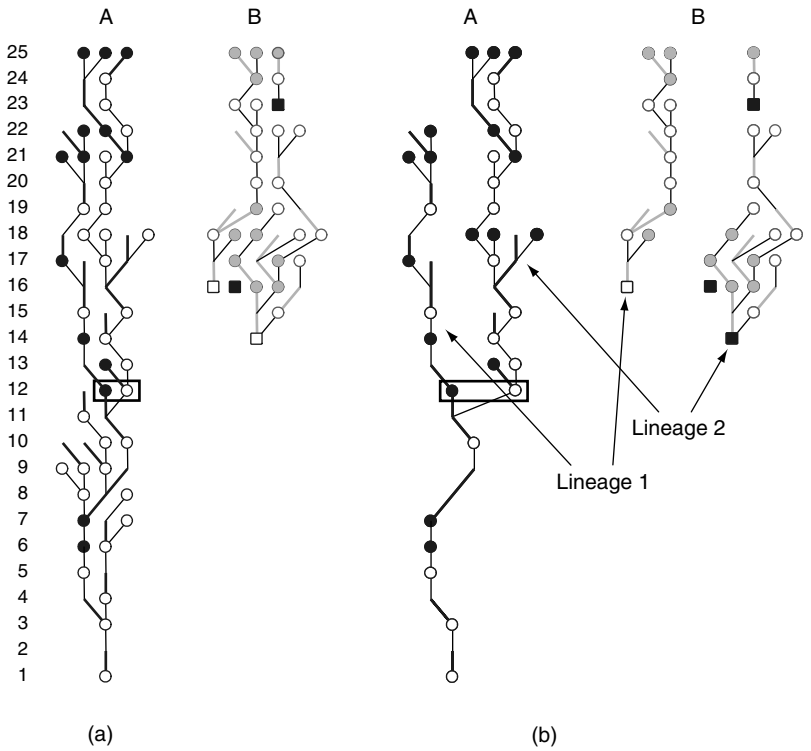


Figure 6.1. A single-species simulation run with unrestricted migration between two continents. (a) The complete phylogeny; (b) the two distinct lineages leading to current species. The boxed species highlight the point of divergence of these lineages.

character model is chosen to be 8 characters and 8 distinct states with a 20% chance of change, because, in reality, the effect must be at least this important.

The results of an initial single run are shown in Figure 6.1. There is quite an ancient divergence, as indicated by the boxed species at the point of divergence (generation 12 out of 25). On continent *B* a recent migration (only two generations back) is seen, leading to the situation where one current species on that continent is only distantly related to the other two current species, themselves lying on an already well-established lineage for this continent. This is apparent in Figure 6.1b, where the two distinct lineages are shown.

The reconstruction for this run is shown in Figure 6.2. The divergent lineages leading to the current species are missed completely in the reconstruction: whereas the true common ancestor is 14 generations back at generation 11 and the true fossil common ancestor (which, ideally, the

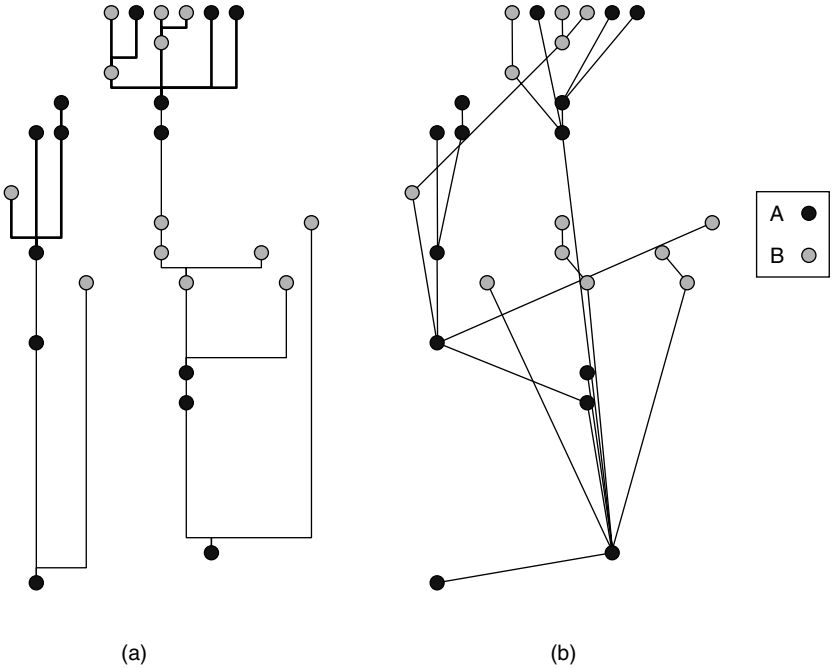


Figure 6.2. Reconstruction based on the simulation in Figure 6.1. Part (a) shows the reconstruction, and (b) the true fossil connections.

reconstruction should identify) is 18 generations back at generation 7, the reconstruction places the common ancestor only three generations back at generation 22! The error results from the two species on continent *B* that are not on the recently migrated lineage. These two species have characters

0110 1001 0011 0001 0001 0000 1000 0000

and

0110 1001 0001 0001 0001 0000 1000 0000,

and although they are on a lineage that split many generations back, they are erroneously connected by the reconstruction to the fossil at generation 24 with characters

0110 1001 0100 0001 0000 0001 1100 0000,

despite differing in five and six characters, respectively. As was found repeatedly in the results presented in the previous chapter, a correspondence in the non-hereditary characters (the first eight characters in this example) leads to incorrect association of fossils. These two should actually connect to the fossil species, also on continent *B*, at generation 19 with characters

0111 1010 0111 0001 0001 0001 0000 0000,

which in turn connects to the species on continent *A* at generation 14, with characters

0111 1010 0100 0000 0001 0001 0000 0100.

The true common ancestor, at generation 11, has characters

0110 0001 0100 0000 0001 0001 0000 0100

and occurs below the boxed species that indicate the point of divergence of these lineages in the figure.

The chance of migration was set to 5% both ways, but owing to the low number of species in this simulation only four migrations occurred, each from *A* to *B*. Specifically, there was one at generation 14, two at generation 16, and one at generation 23. The reconstruction found many more migrations: from *A* to *B* at generations 11 and 13, two at 15, 18 and two at 23, as well as migrations from *B* to *A* at generations 20 and 24. The average time for reconstructed migrations from *A* to *B* was 17, in agreement with the correct value. The back-migrations from *B* to *A* averaged generation 22, so the overall average migration generation in the reconstruction was only slightly later, at generation 18. In summary, there is much more migratory activity according to the reconstruction than actually occurred, and it is placed, on average, slightly later in the evolution.

As in Chapter 5, the average over 1000 runs is used to give a balanced and more accurate overall picture of the simulation when running with these parameters. Figure 6.3 shows that the resulting diversity profile on continent *B* is basically just a scaled-down copy of that on continent *A*. Average values concerning the time and location of common ancestors, overall and for each continent separately, are in presented in Table 6.1.

Because only continent *A* has an initial species in these simulations, there is always a true common ancestor species. This is true overall, and for each continent individually, but for a small percentage of the runs all species on one

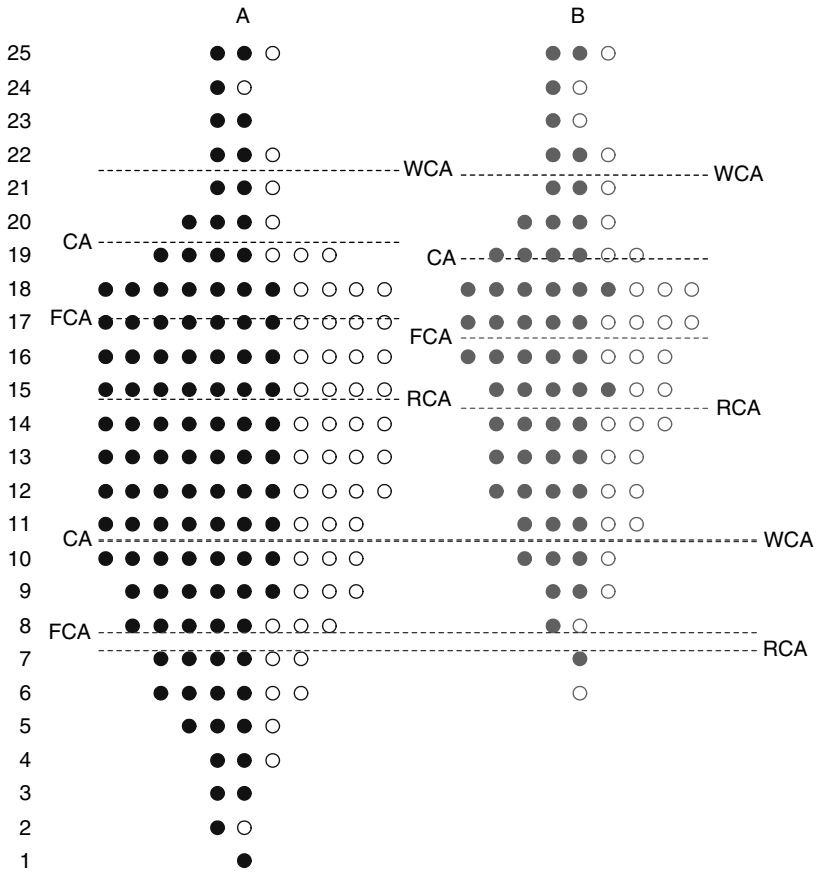


Figure 6.3. Average profiles for 1000 runs of the two-continent amphora migration simulation.

of the continents may vanish. In these cases, the individual continent result will report no true common ancestor, and this is the reason for the non-zero result in the relevant cells of the above table. Runs where species on both continents die out are rejected. The 25 generations usually provide enough time for reciprocal monophyly to develop (Avise, 2000), i.e. all lineages on each continent converge to a common ancestor on the same continent. The degree to which this occurs is substantially underestimated by the reconstruction; the true common ancestor for current species on continent *B* is over six times more likely to occur on continent *B* than continent *A*, but, according to the

Table 6.1. *Average common-ancestor results for the amphora profile with migration*

Note that rounding errors occasionally result in the location percentages not summing to 100%.

			Location		
Ancestor	Continents	Generation	*A*	*B*	none
CA	All	10.6	78%	22%	0%
	A only	19.4	92%	2%	6%
	B only	18.9	13%	79%	7%
FCA	All	7.9	45%	19%	36%
	A only	17.2	79%	9%	12%
	B only	16.6	17%	69%	14%
RCA	All	7.3	58%	17%	25%
	A only	14.8	73%	11%	17%
	B only	14.5	29%	56%	15%
WCA	All	10.6	—	—	—
	A only	21.6	—	—	—
	B only	21.4	—	—	—

reconstruction, it is just less than twice as likely. Even the fossil common ancestor for continent *B* is more than four times more likely to occur on *B* than on *A*, so the reconstruction fails not simply because of the incomplete fossil picture. The situation for current species on continent *A* is similar in some respects, but the performance of the reconstruction is not so different from the true fossil picture. However, both underestimate the degree to which the true most recent common ancestor of the current species on continent *A* actually lies on continent *A*: a massive 98% of the time.

The pattern in the common-ancestor generations is consistent across all methods, with the common ancestor for continents *A* and *B* occurring at similar times and both much more recent than the overall common ancestor generation. However, the actual timing varies substantially. The high fossilisation rate sees the fossil common ancestor values quite close to the true values, but the reconstruction pushes the times further back. Perhaps most surprising is the difference between the true common ancestor and the Wagner common ancestor for the continents alone. Despite close agreement for the overall value, and the results from Section 5.1 where, on average, the Wagner common ancestor was further back in time than the common ancestor for an amphora profile, here the true generation is more than twice as far back as the Wagner-estimated generation for both continents. This is almost certainly a result of the smaller number of current species on each

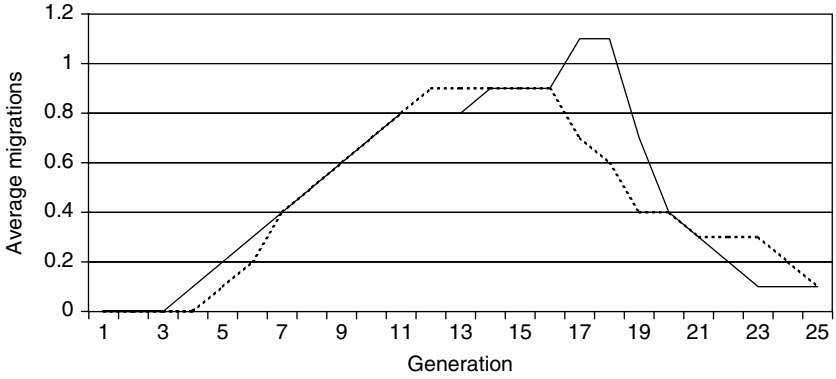


Figure 6.4. True (solid lines) vs. reconstructed (dotted lines) migration times for the amphora average.

continent: an average of 2.7 on each here, compared to 4.5 in the earlier, no-migration result. Certainly, the addition of migration leads to a much earlier generation for the most recent common ancestor, seven species generations further back in fact, because the evolution on each continent is essentially independent. This result alone highlights the importance of an accurate picture of migration when determining times of common ancestry.

Contrary to the single-run result shown in Figures 6.1 and 6.2, the reconstruction does not seem to have a serious problem with the introduction of spurious back-migrations from *B* to *A*. The reconstruction found an average of 7.3 migrations from *A* to *B* and 3.9 migrations from *B* to *A*. This compares very favourably with the average number of true migrations from *A* to *B* of 8.0, and from *B* to *A* of 4.1.

Figure 6.4 shows the number of both true and reconstructed migrations as a function of generation. (For the reconstruction, migration time is approximated by averaging the start and end generations.) Not surprisingly, the time of the peak in migrations corresponds to that of the peak species population. The reconstruction tends to average this peak out somewhat, but nevertheless the true and reconstructed curves do agree quite closely. The average generation of the common ancestor on each continent also corresponds to this migration and population peak period.

The fossil record alone reveals on average only 5.6 migrations from *A* to *B* and 2.0 migrations from *B* to *A*. Although the peak in true migrations is captured, as is generally the case for the fossils-only migration picture, early migrations are missed and the degree of recent migration is exaggerated.

6.2 Simulating hominoid migrations

A more complex species migration scenario is one that broadly corresponds to the pattern evident in the hominoid fossil record. Specifically, continent *A* is the continent of initial population, and corresponds to Africa. Then continent *B* corresponds to Europe, and there is a period of 10 million years where migration from *A* to *B* is possible: specifically, from generations 5 to 15 there is a 10% chance of such a migration. At the same time there is a 5% chance of a back-migration from *B* to *A*. After generation 10, continent *C*, corresponding to Asia, comes into play as migration between *B* and *C* becomes possible for 5 million years. Specifically, from generation 10 to 15 there is a 10% chance of migration from *B* to *C* and a 5% chance of back-migration from *C* to *B*. At generation 17, i.e. after these periods of migration, all species remaining on continent *B* become extinct and thus there is no longer any possibility of passage between continents *A* and *C*. All other settings are as for the amphora profile presented in the previous section.

The results of a single simulation run are shown in Figure 6.5. Historically, most species have been on continent *A*, i.e. the continent continuously populated, although for the most recent several generations continent *C* has had a greater number of species. Similarly, continent *B* has a greater historical species population than continent *C*, although there have been no species on continent *B* for the past 8 million years. For this run fossils are evenly distributed, with five on each continent.

The common ancestor for species on continent *A* is very recent, but there is a deep split in the lineage leading to living species on continent *C*, dating right back to the time of the initial population of this continent. As a result, there is much greater character diversity among living species on continent *C* than on *A* (average of 7.8 cf. 2.0). The overall most recent common ancestor is a new migrant on continent *B* at generation 9, as indicated in Figure 6.5.

Figure 6.6 shows the fossil reconstruction for this run. Of the 15 fossils, only one fossil on *A* and one on *C* are truly on extant lineages, but the reconstruction makes many additional connections and manages an overall accuracy of only 20%. This increases to 62% for current species only, as all three current species on continent *A*, and two of the five current species on continent *B* are correctly associated with their respective parent fossils. The success on continent *A* is clearly a result of the recent common ancestor, and similarly the difficulty with continent *C* is because of deep split in the lineage, and the fact that four of the five fossils lie on extinct lineages. Considering the potentially very important fossils from contintent *B*, the reconstruction performs well for connections not involving migrations, but the others are in error.

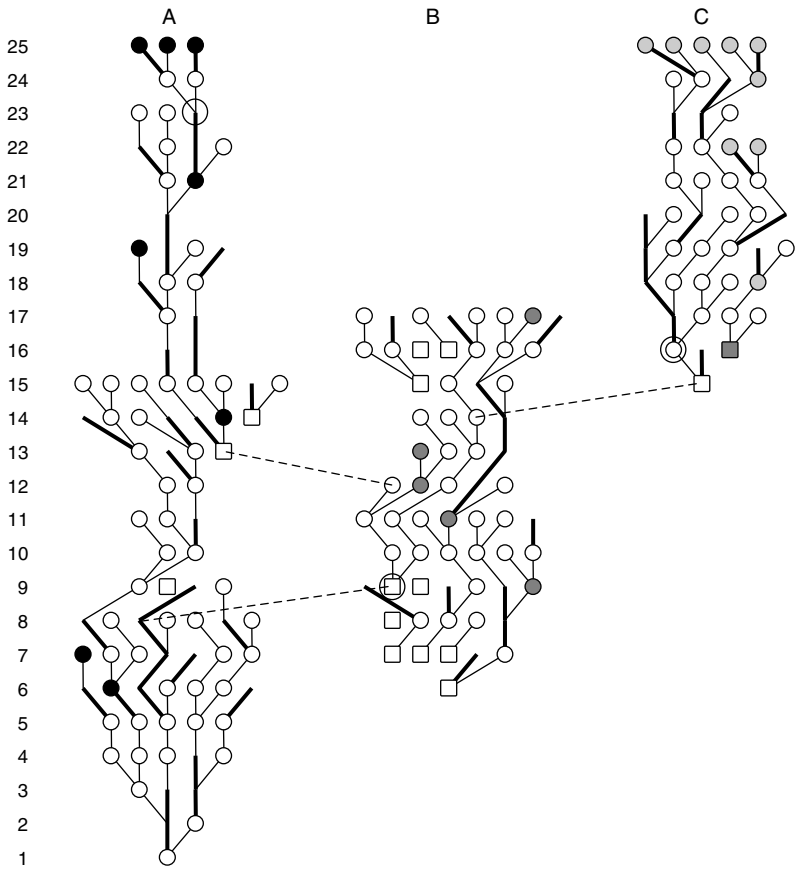


Figure 6.5. Results of a single hominoid migration simulation. The circled species on continents *A* and *C* are the common ancestors for those continents, and the circled species on continent *B* is the overall most recent common ancestor. The two migrations from continent *B* that involve extant lineages on continents *A* and *C* are indicated by the dashed lines, as is the migration from *A* to *B* that involves the overall common ancestor. The many other migrations are apparent from the number of species drawn as squares.

Clearly in this scenario, determining details of the migrations is fundamentally important, but of the 15 migrations that occurred, none is represented in the fossil record! This is clear from Figure 6.6b, where no fossils from different continents are connected. However, the reconstruction finds five migrations: three from *A* to *B* starting after generations 6 and 7, and one each from *B* to *C* and *B* to *A*, occurring much later, at generations 12 and

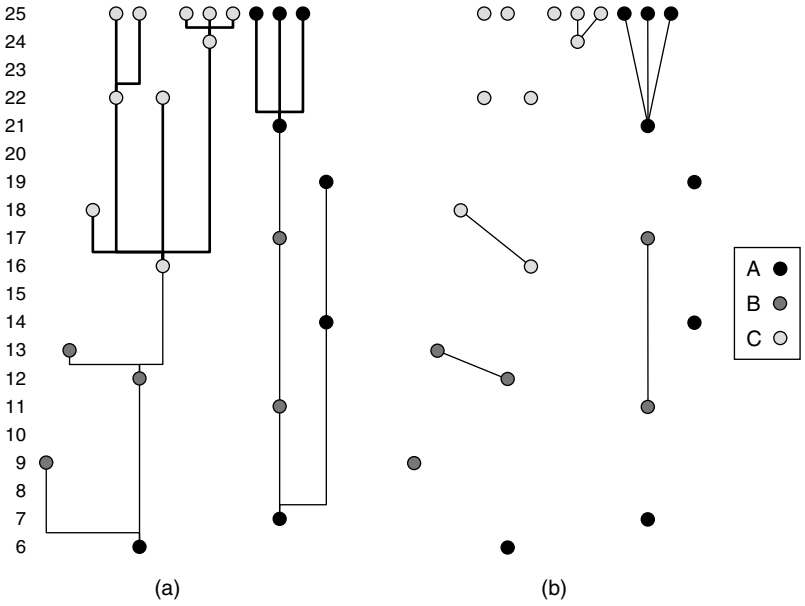


Figure 6.6. (a) The reconstructed phylogeny for the single hominoid simulation; (b) the true fossil phylogeny. Most notable are the many spurious connections identified by the reconstruction in all cases except the phylogeny of recent species from continent *A*.

17 respectively. The major migratory activity between generations 6 and 9 is thus not particularly reflected in the reconstruction, and that from generations 13 to 16 is almost not apparent at all.

As before, averaging over 1000 runs provides a more robust view of the situation under this scenario. The resulting profiles and common ancestor generations are illustrated in Figure 6.7.

Essentially the same pairwise difference across current species on continents *A* and *C* is seen, unlike in the example single run above, and despite the continent *C* species population being much younger. Both continent *A* and continent *C* have quite a recent common ancestor, at approximately generation 20, but the overall common ancestor is well back on average, in fact prior to the start of any migrations, between generations 4 and 5.

The true overall common ancestor is nearly always on continent *A*, although for 7% of runs the overall common ancestor is found on continent *B*, and, indeed, in one run from this set of 1000, the common ancestor is on continent *C*. The common ancestor for continent *C* alone is found on continent *A* 4% of the time and on continent *B* a little over 1% of the time; otherwise it is

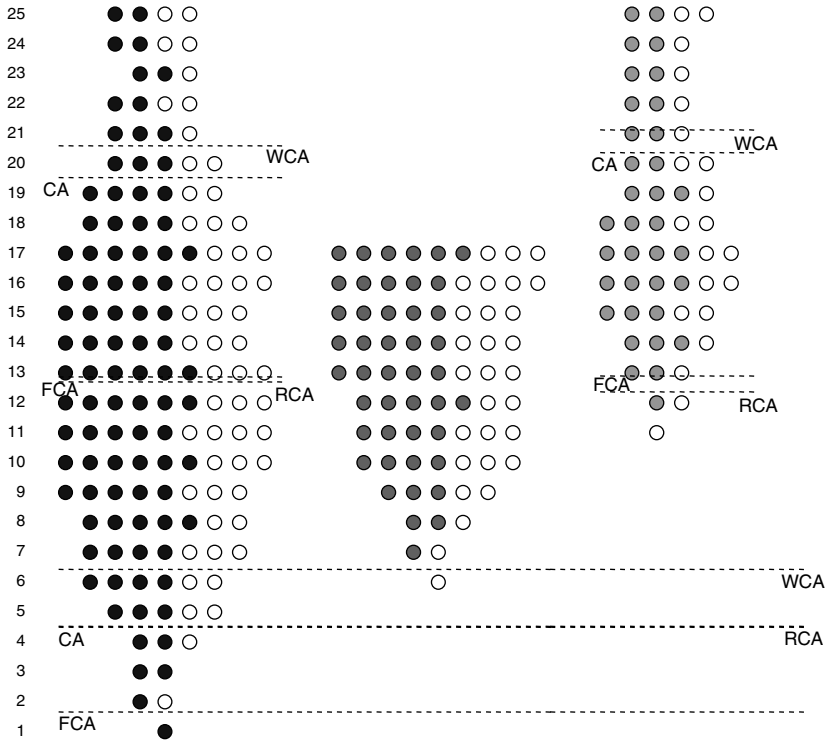


Figure 6.7. Average hominoid migration simulation.

on continent *C*. The common ancestor for the current species on continent *A* was found on continent *A* in all cases.

A true overall fossil common ancestor is relatively rare, occurring in approximately 22% of runs, with nearly 90% of these on continent *A*. However, this was not the case for the individual continents, where a fossil common ancestor was found in more than 80% of the runs. In continent *A* alone, the fossil common ancestor was nearly always also on continent *A*, but for continent *C*, although in the majority of cases the common ancestor was also on continent *C*, there remained a significant possibility of its occurring on either of the other two continents.

In the individual continent cases, there is fairly close agreement between the true fossil common ancestor locations and those obtained from the reconstruction, with just a slight overestimation of the frequencies in the reconstruction case for continent *A*, and some spurious assignment to continent *A* in the continent *C* case. However, for the overall reconstruction common ancestor,

an ancestor is identified in nearly two thirds of the runs, and subsequently the figures are substantially inflated for each continent.

Only 20% of fossils lie on surviving lineages and, as has been found consistently in the species simulations, the reconstruction places approximately double this number: 39% in this case. Of particular interest is the fact that, on average, only 6% of fossils on continent *B* lead to current species (i.e. 0.4 of 7.0) despite the significant number of species on this continent, and their crucial role in the linking of continents *A* and *C*. The actual connection accuracy for the fossil reconstruction is only in the 30% range, rising to roughly 50% for current species.

According to the reconstruction there are a significant number of migrations directly between continents *A* and *C*, although such migrations are excluded by the settings of the simulation. However, from a fossil perspective, such migrations are possible because the intermediate, continent *B* resident species may not be fossilised. For this reason, it is interesting to compare the migrations according to the reconstruction with both all true migrations and those migrations evident from the fossil record alone. In this case, when comparing with true migrations as mentioned above, the reconstruction identifies disallowed migrations between continents *A* and *C*, and underestimates the other migrations by factor of around 2 or 3. However, when compared with the record of migrations from fossils alone, the reconstruction identifies significant extra migrations in all cases. It is worth mentioning at this point that the fossil record does not necessarily reduce the number of migrations apparent. For example, if a single migrant species remains unfossilised yet its immediate descendants (that reside on the destination continent of course) are fossilised, the fossil record will indicate a distinct migration for each of these descendant species, as opposed to the single true migration. Such issues need to be kept in mind when interpreting any reconstructed phylogeny, because the fossil reconstruction process employed here (unlike the Wagner reconstruction process) does not introduce hypothetical species.

The identification of the clades including current species is quite successful for these simulations, managing approximately 80% monophyly. However, the Wagner reconstruction is much less successful in this regard, with the degree of monophyly only 40%.

6.3 Restricted migrations and interbreeding

In this section, the results of some more generic migration simulations are presented, to study the effects of general restrictions in migration. Firstly, the

case where there is just a small window of time in which migration is allowed is studied; the aim is to quantify any differences in the results as this window moves. This corresponds to the physical situation where barriers form and disappear, perhaps owing to particular geographic events or formations, or to continental drift. Secondly, various source–sink scenarios are simulated, where migration is not limited in time, but rather in direction; one continent acts as the source of all migrations, and two others as sinks. This corresponds to a physical situation where perhaps ocean currents, or particular geographic boundaries, effectively create one-way passages between otherwise isolated regions. Of particular interest in these simulations is the effect of differing selective advantage across continents.

A logistic profile is used in each case, with substantial interbreeding: specifically, a 50% likelihood of lineage merging if the two taxa differ in three or fewer characters and share a common ancestor in the last three generations. The fossilisation rate is kept at a constant 10%. The reconstruction does quite well with this level of fossilisation, and thus many reconstruction difficulties associated with a lack of fossils, the effects of which are described in Section 5.3.1 and fairly well understood, may be removed. Nevertheless, the rate remains not so large as to make the results inapplicable to real situations. The non-hereditary character parameters remain at their default setting of 8 characters and 8 distinct states with a 20% chance of change.

6.3.1 Migrations restricted in time

The simulations presented in this section involve migrations between two continents, but the migrations are restricted to a small window of six generations only, at different times in the evolution: specifically between generations 5 and 10 (*early*), 10 and 15 (*middle*), 15 and 20 (*late*) or 20 and 25 (*current*).

Figures 6.8 and 6.9 show the results from two simulations, each with migrations between continents *A* and *B* occurring with a probability of 5%, but restricted in time as described above. Specifically, Figure 6.9 shows a run with the *late* window, i.e. the migration is restricted to between generation 15 and generation 20, and in Figure 6.8 the *middle* window is used, restricting migrations to the six generations from generation 10 to generation 15. The primary difference between the two cases is the fact that, despite having a very early overall common ancestor in both cases (at generations 5 and 6 respectively), Figure 6.8 shows that for the earlier migration case, there has been enough time for one lineage on continent *B* to become dominant and ensure that the current taxa there come from purely continent-*B* lineages (since the most recent common ancestor). However, reciprocal monophyly

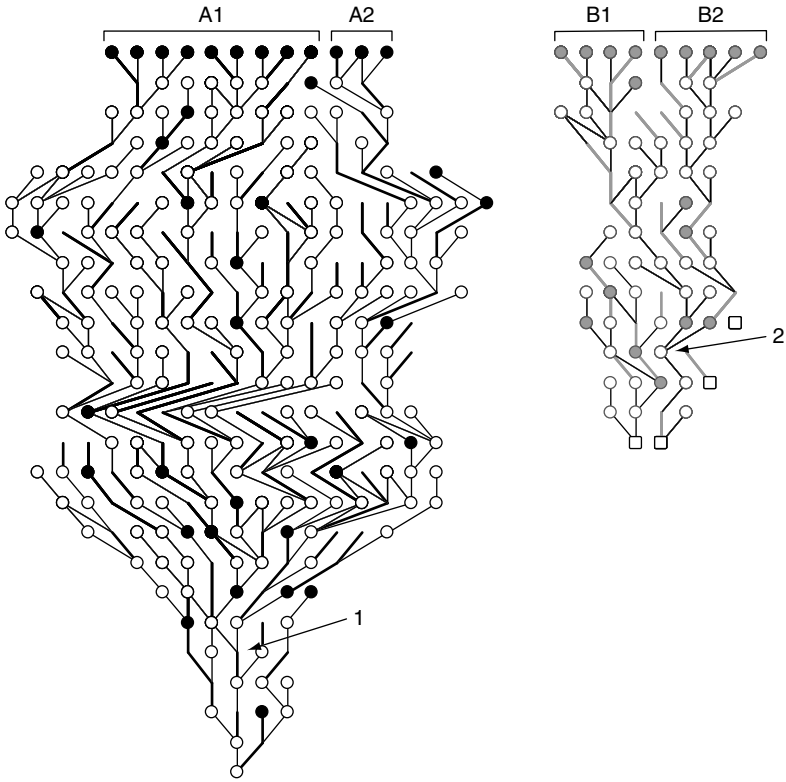


Figure 6.8. Results of a single run with migration only possible between generations 10 and 15. The overall most recent common ancestor is the species indicated by the arrow labelled 1, and this species is also the last common ancestor for continent *A* alone. Similarly, the continent-*B* most recent common ancestor is indicated by the arrow labelled 2. The two surviving lineages after the divergence at each of these points are illustrated by the brackets at the top of the figure. The lineage leading to A1 also includes the continent-*B* common ancestor.

(Avise, 2000), as discussed in the Section 6.1, has not developed, because the most recent common ancestor for the current species on continent *A* is also a common ancestor for the current taxa on continent *B*. In the later migration case, there has not been sufficient time for any lineage-sorting to be completed.

Figure 6.10 shows the reconstruction and clade construction for the *middle* window run in Figure 6.8. All four of the clades that contain current species are 100% correct, and all contain species from a single continent only. The three earlier clades are polyphyletic, finding all true ancestors of the parent species in each case, but also including extra species. Some of the current

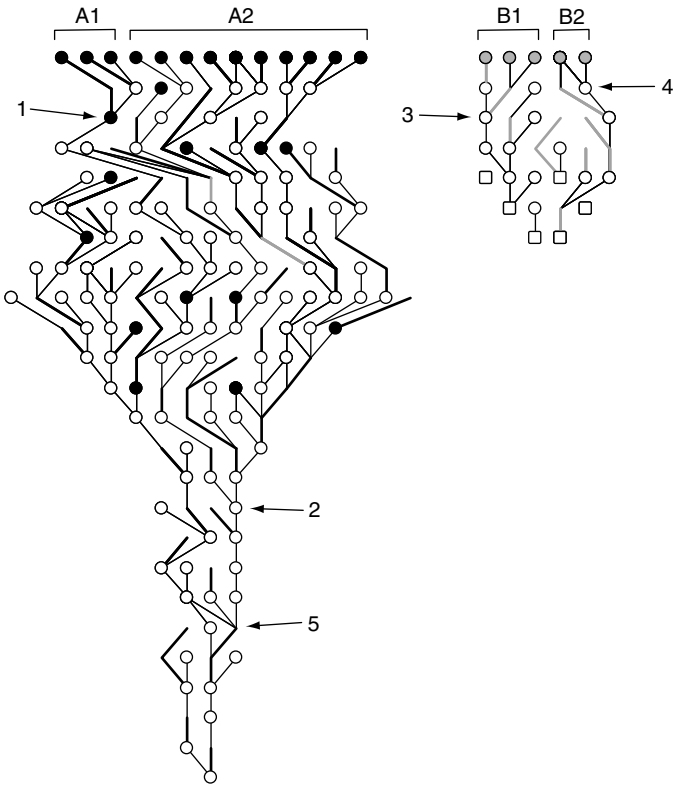


Figure 6.9. Results of a single run with migration only possible between generations 15 and 20. Several species of interest are labelled: species 1 is the most recent common ancestor (CA) of the current species under the bracket labelled A1; species 2 is the CA of the species labelled A2; species 3 is the CA of the species labelled B1; species 4 is the CA of the species labelled B2; and species 5 is the overall most recent common ancestor. Species 2 is also ancestral to the group of species labelled B2, and this species is also the last common ancestor for continent *A* alone. Similarly, species 2 is the continent-*B* most recent common ancestor. The two surviving lineages after the divergence at each of these species are illustrated by the brackets at the top of the figure.

species remained unallocated to a clade, because the potential clade would be too inclusive and thus violate the 'clade size' limitation used in this particular reconstruction (see Algorithm 2 in Section 3.1.1, and associated discussion). The overall current species reconstruction accuracy was 85% for this simulation, but when considering all fossils this dropped to only 44%. As usual, the reconstruction overestimated the number of fossils on surviving

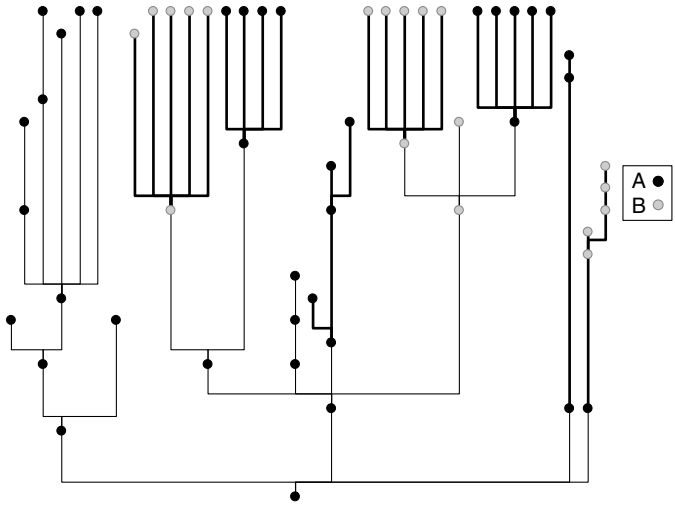


Figure 6.10. Fossil reconstruction for the simulation shown in Figure 6.8.

lineages, in this case by 50% with 12 instead of 8 of the 36 fossils identified as being ancestral to current species.

Figure 6.11 shows the average results after 1000 simulations for all four positions of the migration window: *early* (a), *middle* (b), *late* (c), and *current* (d). Important diversity values are summarised in Table 6.2. The early and middle migration windows allow enough time for equilibrium in the number of taxa to be established on the destination continent (recall that the defining feature of the logistic profile is growth to an equilibrium value). For the late migration window it is close, but equilibrium on continent *B* is not quite established. In the current window case, the diversity has not reached equilibrium on either continent. The total number of taxa on continent *A* is maximum in this case, because there has only been minimal loss of species to continent *B*.

The pairwise character difference between current taxa shows a particularly interesting trend for continent *B*. Overall, there is very little change in this value across the four scenarios: the value ranges only from 11.1 for the current window to 11.4 for the early and middle windows. Contrary to this, the value for continent *A* alone decreases from 10.9 to 10.2 as the migration window moves further back in time. The value for continent *B* also decreases, but much more markedly with the values of 9.5, 8.5, 8.1 and 7.3 for current, late, middle and early migration windows respectively: a 23% drop. This gives a clear indication of lineage sorting. Overall, divergent lineages can persist on

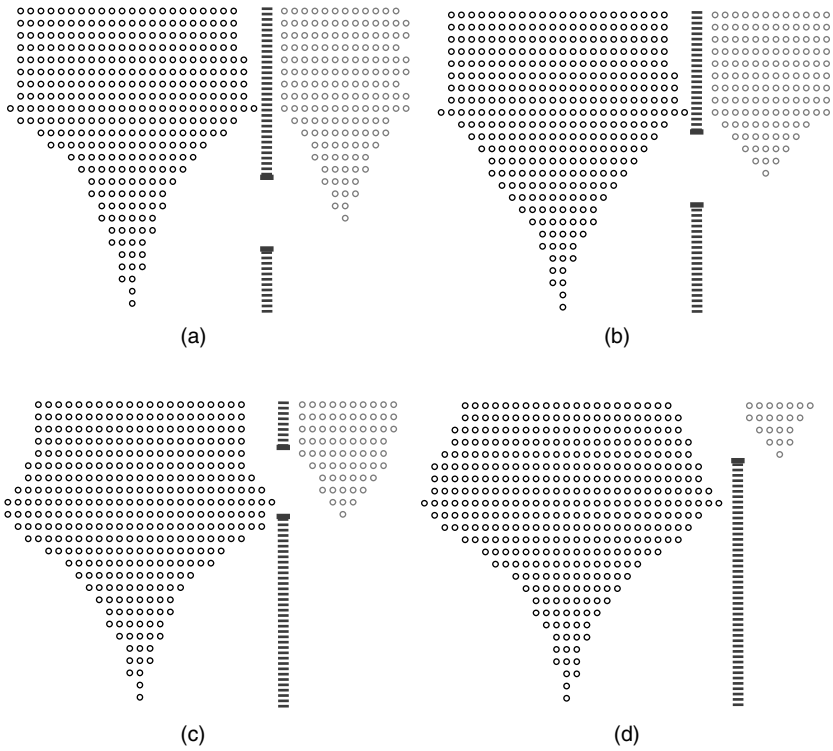


Figure 6.11. Average profiles for four different migration periods, with the barriers restricting migration in each case as shown.

the different continents, but on any one continent, as time goes by, particular lineages will begin to dominate and thus diversity will decrease.

Related to this is the nature of the common ancestry overall, and on each continent separately, as summarised in Table 6.3. It can be seen from these results that for later migration, current taxa on continent *B* are more likely to be monophyletic, as found in the single run examples in Figures 6.8 and 6.9. Parallel to the increase in current taxa character differences as the migration window moves later (discussed above and shown in Table 6.2), the later the migration window, the earlier the common ancestor for the individual continents. At the same time, the overall common ancestor moves slightly earlier in time: from generation 5 back to just later than generation 4.

The Wagner common-ancestor estimate is consistently incorrect, overall and for each continent individually, highlighting the continued inability of the Wagner reconstruction to cope with recent maximum profiles,

Table 6.2. *Total number of taxa, number of current taxa, and average pairwise character differences between current taxa for the barrier migration runs*

	Total	*A*	*B*
Early			
Total	331.5	222.8	108.6
Last	35.4	22.5	13.0
Differences	11.4	10.2	7.3
Middle			
Total	320.5	231.4	89.1
Last	35.2	22.3	12.9
Differences	11.4	10.3	8.1
Late			
Total	292.2	243.6	48.6
Last	32.3	21.7	10.5
Differences	11.2	10.5	8.5
Current			
Total	274.6	259.8	14.7
Last	29.0	21.3	7.6
Differences	11.1	10.9	9.5

interbreeding and non-hereditary characters. This is especially interesting in the continent-*B*-only case because there are very few taxa, and the true common ancestor is not right back at the start of the simulation as in the logistic profile cases studied in Section 5.2, thus removing two significant complicating factors for this algorithm.

The fossil common-ancestor values echo the true common-ancestor ones, but occur much earlier in the simulation because of sampling. The differences are not so pronounced because a fossil common ancestor does not exist in so many cases: always at least 50% and often considerably more. The reconstruction performs very well, with a common-ancestor generation no more than two generations different from the true fossil common-ancestor generation in any case. However, a reconstruction common ancestor is found significantly more often (roughly 1.5–2 times more often).

Average values comparing real and reconstructed migrations are given in Table 6.4. There are very few true migrations in the early migration window case owing to the low number of taxa while the window is open. The number of migrations correspondingly increases as the migration window moves later in the simulation and is maximum when it coincides with the

Table 6.3. *Time and location of common ancestors for the barrier migration runs*

The difference between 100% and the sum of the location percentages in brackets gives the percentage of runs where no ancestor existed (or was identified) for the particular case.

	generation (location: *A*% – *B*%)			
	Early	*Middle*	*Late*	*Current*
CA				
All	4.2 (99% – 1%)	4.7 (99.7% – 0.3%)	4.9 (99.7% – 0.3%)	5.0 (100% – 0%)
A only	6.9 (99% – 0%)	6.7 (99% – 0%)	6.1 (99% – 0%)	5.4 (100% – 0%)
B only	9.1 (29% – 53%)	10.5 (54% – 41%)	9.9 (71% – 27%)	7.7 (90% – 9%)
WCA				
All	−2.4 (–)	−2.3 (–)	−1.7 (–)	−1.1 (–)
A only	8.1 (–)	8.6 (–)	8.0 (–)	6.3 (–)
B only	16.5 (–)	16.6 (–)	17.7 (–)	19.1 (–)
FCA				
All	2.0 (32% – 0.4%)	2.1 (34% – 0.2%)	2.3 (37% – 0.2%)	2.3 (36% – 0%)
A only	3.5 (44% – 0%)	3.6 (42% – 0%)	3.2 (41% – 0%)	2.5 (38% – 0%)
B only	5.5 (21% – 26%)	6.7 (34% – 23%)	6.0 (44% – 11%)	4.4 (44% – 4%)
RCA				
All	3.1 (57% – 1%)	3.4 (60% – 0.4%)	3.3 (61% – 0.3%)	3.4 (64% – 0%)
A only	4.3 (62% – 1%)	4.5 (64% – 0.4%)	4.1 (63% – 0.4%)	3.7 (65% – 0.1%)
B only	5.2 (39% – 18%)	6.6 (53% – 16%)	6.2 (62% – 9%)	5.9 (71% – 3%)

diversity maximum. However, the fossil reconstruction migrations peak for the middle window rather than the late window, because of the reconstruction's general tendency to overestimate the number of migrations and also to spread them over a larger time window. As a result, although the true number of migrations is 50% higher for the late window than for the middle window, the reconstruction determines an almost equal number of $A \rightarrow B$ migrations, and 50% more $B \rightarrow A$ migrations with the middle window. In general, the number of migrations is overestimated significantly by the reconstruction for the earlier migration windows, but matches more closely for the later ones. In all cases several spurious back-migrations are found: 35% of all migrations for the early case, but gradually getting less important as the migration window moves later. This occurs even though the reconstruction algorithm favours fossil connections that do not cross continents.

A more telling picture is provided by the curves in Figure 6.12. The reconstructed migrations exhibit much more pronounced spreading when compared with the amphora-profile migration simulation presented in Section 6.1.

Table 6.4. *True and reconstructed migrations for the barrier migration runs*

	Early	*Middle*	*Late*	*Current*
Real $A \rightarrow B$	2.2	5.8	8.5	6.6
Reconstructed $A \rightarrow B$	5.4	8.1	8.3	7.2
Reconstructed $B \rightarrow A$	2.9	3.3	2.1	0.7
Total reconstructed	8.3	11.4	10.4	7.9

Indeed, for the early migration window, there is not even any discernible peak in the migration distribution: the reconstruction in this situation has not managed to determine any restriction on the migrations at all. The area under each curve corresponds to the total number of reconstructed migrations, as reported in Table 6.4.

An interesting comparison is achieved by looking more closely at the simulation with the early migration window, and seeing how it developed during the course of its evolution by taking snapshots at generations 11, 16 and 21. These snapshots may then be compared directly with the above results for the current, late and middle migration windows respectively. Table 6.5 shows the common ancestor times and locations for snapshots at these three generations, as well as the final values for the early migration window results above. As is to be expected, the true common ancestor generation increases as the evolution proceeds further, owing to lineage-sorting. This effect is most pronounced for continent *B*, owing to its lower overall population, but is nevertheless clearly evident both overall and for continent *A* alone. The continent-*B* common ancestor also becomes increasingly more likely to be located on continent *B*, as indicated by the chart in Figure 6.13a. However, although this variation is also evident in the fossils, the degree to which it is captured by the reconstruction is minimal. According to the reconstruction, the likelihood of a continent-*A* or a continent-*B* location for the continent-*B* common ancestor changes very little, especially across the first three snapshots. The Wagner common ancestor calculation once again shows little agreement with the true values.

The plot in Figure 6.13b shows that the average pairwise character difference between taxa across both continents is only marginally greater than that on continent *A* alone. This is because the same period of evolution is involved, and the small number of taxa and shorter duration of the evolution on continent *B* mean that only a small amount of independent character diversity has been maintained there. The other interesting feature is the way in

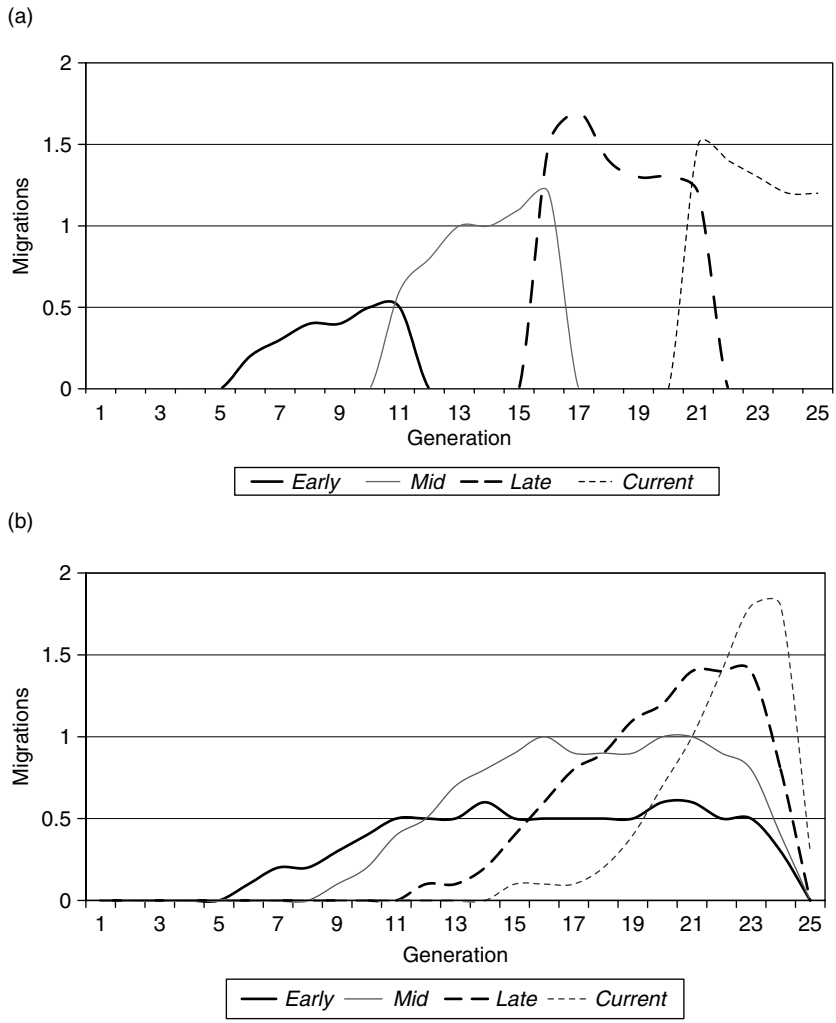


Figure 6.12. True (a) and reconstructed (b) migration times for the barrier migration runs.

which the curves are flattening out, indicating a significant slowing in the generation of diversity, parallel to the population saturation inherent in the logistic profile.

Yet another interesting comparison is with a simulation where, rather than a high rate of migration restricted to a small time window, there is

Table 6.5. *Time and location of common ancestors for the snapshots at generations 11, 16, 21 and 25 for the early migration run*

The difference between 100% and the sum of the location percentages in brackets gives the percentage of runs where no ancestor existed (or was identified) for the particular case.

	generation (location: *A*% – *B*%)			
	11	16	21	25
CA				
All	2.8 (100% – 0%)	3.6 (99.7% – 0.3%)	3.9 (99% – 1%)	4.2 (99% – 1%)
A only	3.3 (100% – 0%)	4.7 (99.5% – 0%)	5.7 (99% – 0%)	6.9 (99% – 0%)
B only	4.7 (67% – 25%)	6.6 (45% – 39%)	7.5 (37% – 47%)	9.1 (29% – 53%)
WCA				
All	1.1 (–)	−0.4 (–)	−1.6 (–)	−2.4 (–)
A only	4.4 (–)	5.3 (–)	6.9 (–)	8.1 (–)
B only	8.5 (–)	11.5 (–)	14.0 (–)	16.5 (–)
FCA				
All	1.3 (22.4% – 0%)	1.6 (31% – 0.1%)	1.7 (32% – 0.4%)	2.0 (32% – 0.4%)
A only	1.5 (25% – 0%)	2.1 (38% – 0%)	2.8 (41% – 0%)	3.7 (44% – 0%)
B only	2.7 (26% – 12%)	3.0 (27% – 12%)	4.2 (27% – 20%)	5.5 (21% – 26%)
RCA				
All	2.1 (48% – 0.1%)	2.7 (55% – 0.8%)	2.9 (58% – 2%)	3.1 (57% – 1%)
A only	2.4 (51% – 0.3%)	3.1 (59% – 1%)	3.5 (61% – 2%)	4.3 (62% – 1%)
B only	3.5 (46% – 12%)	3.7 (47% – 10%)	4.3 (44% – 14%)	5.2 (39% – 18%)

a slower rate of migration spread across a large number of generations: as, indeed, the reconstruction infers from the early and middle migration window results.

Allowing migration between continents *A* and *B* from generation 5 to generation 20 with a probability of 2% gives a situation where many of the measures are very similar to the middle window case above. The true number of migrations is almost identical, as are both the number of reconstructed migrations and their generation profile. Further, the overall and continent-*A*-only results are extremely close to their corresponding middle migration window values. The only results that hint that the migration history is actually different in this case are those concerned with continent *B* alone. Firstly, there is a small but significant increase of 7.4% in the character differences among current taxa on continent *B*, due to the more prolonged, and, in particular, later influence of continent-*A* lineages. This prolonged influence of continent *A* on continent *B* also interrupts lineage-sorting, and the common ancestor is consequently further back, 16 generations rather than 14.5. This feature is similarly apparent in the other common ancestor estimates.

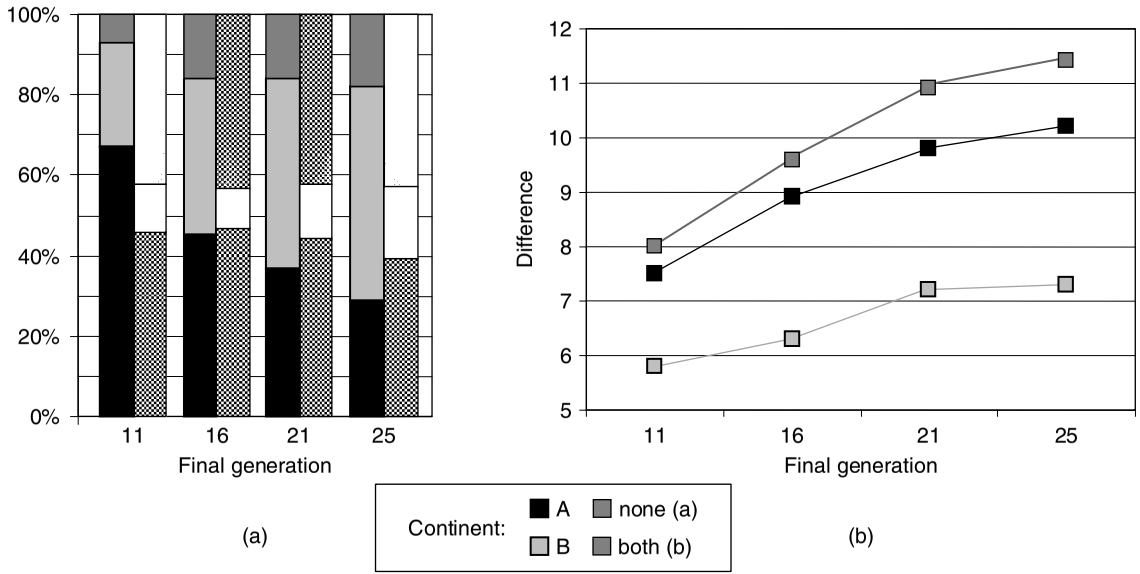


Figure 6.13. Two plots summarising data from the different snap shot times for the early window migration run. (a) The continent-*B* common ancestor location percentages as a function of final generation. The solid bars indicate the true value, and the shaded bars the value according to the reconstruction. (b) The average pairwise character difference between current taxa overall and for each continent alone.

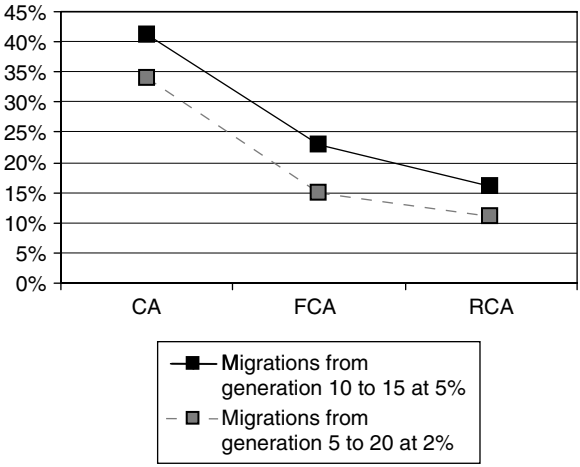


Figure 6.14. This plot shows the likelihood for the three continent-*B* common ancestors, true, fossil and reconstructed, to be located on continent *B* for two different barrier migration simulations. The solid line shows the case where there is a small migration window but a relatively high chance of migration, whereas the broken line shows the opposite case, where there is a longer period of allowed migrations but less chance. When the migratory contact between the continents is prolonged, the chance that the continent-*B* common ancestor actually occurred on continent *A* is also increased.

The most significant indication comes, not surprisingly, in the common ancestor location results. As above, there is no difference between the two simulations for the overall and continent-*A*-only results, but the continent-*B* common ancestor is much less likely to be located on continent *B* when there has been prolonged species flow from continent *A*. This is seen clearly in Figure 6.14.

6.3.2 Migrations restricted in direction

A situation different from that studied in the previous section arises when, rather than keeping the different continents isolated except for a short period of time, there is ongoing migration but only one the continents is allowed to act as a source. This is of particular interest in the case where there is different selective advantage between the continents. The simulations in this section involve three continents, one where only outgoing migrations are allowed (the *source* continent) and the other two allowing only

incoming migrations (*sink* continents). Three specific cases are considered: one where there is no difference in selective advantage between the continents, one where taxa from the source continent have a clear selective advantage, and one where taxa on the destination continents have a clear selective advantage.

Figure 6.15 shows the results of a single run where the sink continents have the selective advantage. (The degree of selective advantage in all the simulations in this section is such that four generations of evolution are required for any advantage to be completely developed or lost in migrant species.) The boxed taxa at the fifth generation of the simulation and the lineages highlighted in the figure show clearly that the populations on each of the sink continents have an ancient separation. Selective advantage on any particular continent implies that established lineages have a better chance of persisting than do lineages resulting from more recent migrations. In this case, each sink continent has acted to preserve a different ancient lineage.

One effect of this preservation is that the common ancestor of all species is on continent *A* very early in the simulation, at generation 4. Reciprocal monophyly has developed on all continents, with the common-ancestor taxon six generations back for continent *A* and ten generations back for both continent *B* and continent *C*. Because of the one-way nature of the migrations, the continent-*A*-only common ancestor is the most recent of the three. However, the fossil common ancestor for continent *B* lies right back at generation 3 on continent *A*! The fossil reconstruction, shown in Figure 6.16, identifies this same fossil as the most recent common ancestor for all three continents, together and independently. The true common ancestor for continent *C* is circled in the figure, but this was not recovered by the reconstruction. One obvious feature of the reconstruction is a deep division, indicated by the dashed line. This is a true feature of the run, not an artefact of the reconstruction: all fossil taxa have been placed by the reconstruction on the correct side of this division.

Table 6.6 shows the diversity and character-difference results averaged over 1000 simulations for the three selective advantage scenarios described above. The total number of taxa is maximum for the case where the source continent has the selective advantage, because in this case taxa from continent *A* are able to thrive on the other two continents after migration. Conversely, total taxon population is a minimum for the case where the sink continents have the advantage, because they produce no migrant taxa, and migrants from continent *A* are less successful. The most significant effect is seen in the figures for the sink continents only. The size of the last generation varies by as much as 18%, and the character difference for current taxa by 47%. This last value is particularly interesting in comparison with the overall and source

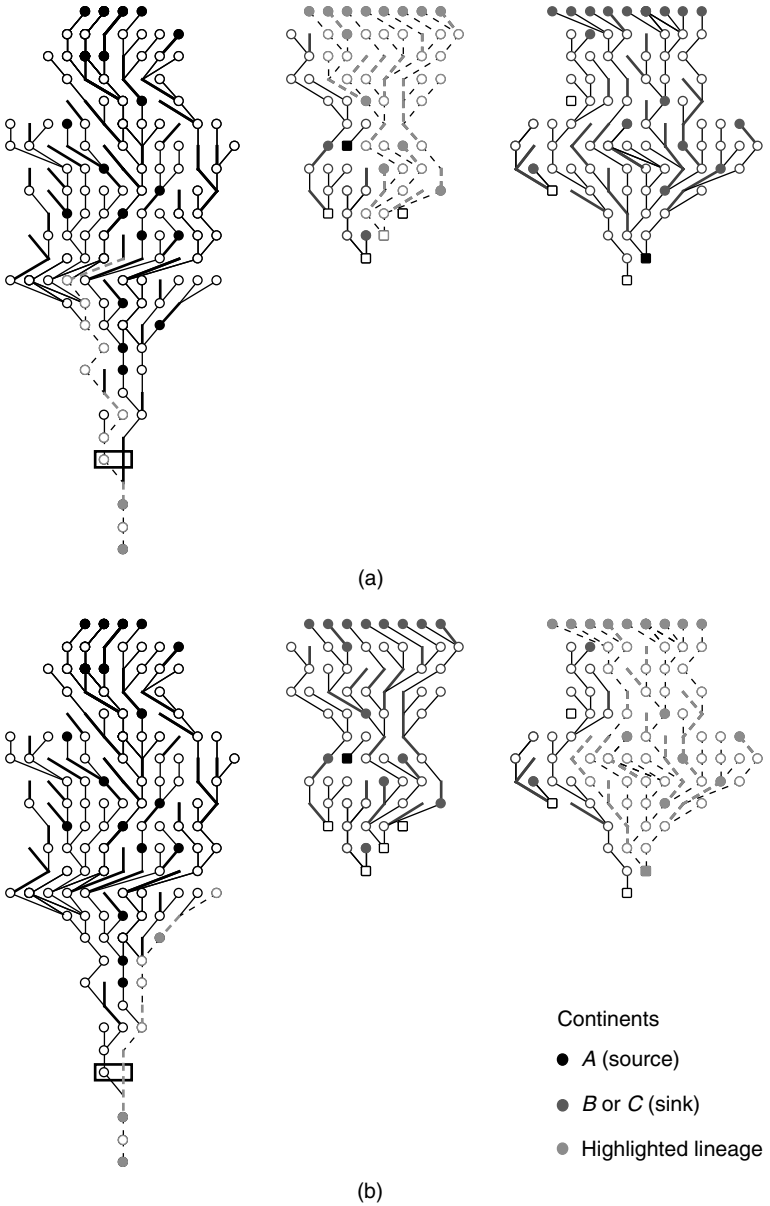


Figure 6.15. Source–sink migration run with sink advantage. The boxed taxa indicate the point of divergence between the lineages leading to extant species on each of the sink continents, as highlighted by the dashed lines. See text for further details.

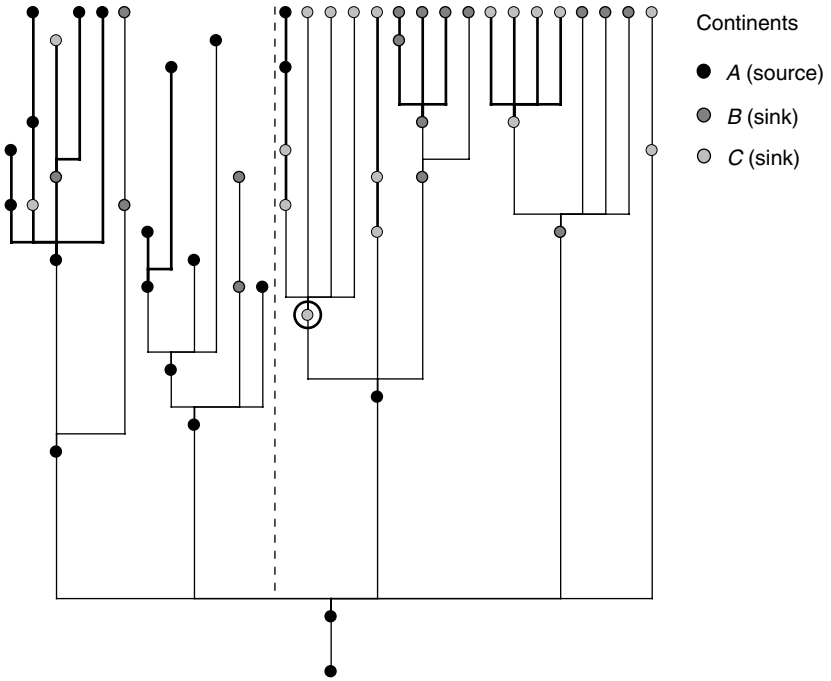


Figure 6.16. Fossil reconstruction for the source–sink migration run with sink advantage. The circled fossil is the true fossil common ancestor for continent *C*, not recovered by the reconstruction. However, the division indicated by the dashed line is a true feature of the run.

continent results, each of which show only a negligible change of close to 1% in average current taxon character difference.

The average common ancestor generation and location results are summarised in Table 6.7. Again, the most significant effect is seen in the sink continent results, both in terms of generation and location. In particular, the percentage of times that a sink continent common ancestor is on the same continent shows a great deal of dependence on the selective advantage, being almost twice as likely when the sink continents are advantaged than when the source continent is. The table only displays the percentage of runs where any particular common ancestor occurs on either sink continent, but, apart from the reconstruction common ancestor, the particular ancestral taxon for one sink continent can never occur on the other, because there is no migration between them. In the fossil and reconstruction results, the generation changes relatively little across the three cases.

Table 6.6. *Total number of taxa, number of current taxa, and average pairwise character differences between current taxa for the source–sink migration runs*

	Total	*A*	*B*	*C*
No advantage				
Total	388	148	119	122
Last	45.6	11.0	17.3	17.3
Differences	11.7	8.3	9.5	9.6
Source advantage				
Total	436	153	135	147
Last	53.1	11.6	20.3	21.2
Differences	11.8	8.3	10.0	10.3
Sink advantage				
Total	366	144	110	112
Last	41.2	10.6	15.0	15.6
Differences	11.7	8.2	9.2	9.2

As for the single-run example shown previously, the continent-*A* current taxa consistently have the most recent common ancestor. This is most pronounced when continent *A* has the selective advantage, because the more successful migrant taxa on the other continents in this case retard lineage sorting.

The number and breakdown of true and reconstructed migrations are shown in Table 6.8. When estimating the number of migrations from continent *A*, the reconstruction is quite close except in the source advantage case, where there is an overestimation of 26%. This is a result of the increased number of taxa, combined with the general habit of the reconstruction to be overly connective. For this same reason, there are many spurious sink-to-source back-migrations in all three cases. The majority of this extra activity occurs late in the evolution, as indicated in Figure 6.17. The profile of the real migrations shows very little difference across the three cases, and the reconstruction profiles are consistently higher than the true curves. The maximum overestimation is for the source advantage case and the minimum for sink advantage case, corresponding to the results shown in Table 6.8.

Both the source–sink simulations described in this section and the barrier migration simulations from the previous section report very high numbers of migrations in the true fossil record: higher than both the true number of migrations and the number of reconstructed migrations. This is a direct consequence of the logistic profile used and the relatively large number of current species (actually subspecies) that result. The true fossil record migration

Table 6.7. *Time and location of common ancestors for the source–sink migration runs*

The figures in brackets show the percentage of runs in which the relevant species occurred on the source continent (*A*, the first figure) or on either sink (*B* or *C*, the second figure).

	generation (location: $A\% - B + C\%$)		
	No advantage	Source advantage	Sink advantage
CA			
All	4.0 (99.3% – 0.7%)	3.9 (99.6% – 0.4%)	4.2 (99% – 1%)
A only	10.4 (95.6% – 0%)	10.5 (96.3% – 0%)	10.7 (96% – 0%)
B only	7.2 (77% – 20%)	6.6 (83% – 15%)	8.2 (71% – 26%)
C only	7.5 (76% – 22%)	6.3 (84% – 14%)	8.1 (71% – 26%)
WCA			
All	−3.7 (–)	−4.3 (–)	−3.1 (–)
A only	19.4 (–)	20.0 (–)	19.1 (–)
B only	15.2 (–)	14.6 (–)	16.1 (–)
C only	15.2 (–)	14.3 (–)	15.6 (–)
FCA			
All	1.8 (31% – 0.3%)	1.9 (31% – 0.2%)	1.9 (32% – 0.6%)
A only	6.4 (56% – 0%)	6.6 (56% – 0%)	6.5 (58% – 0%)
B only	4.2 (34% – 10%)	3.6 (36% – 6%)	4.6 (35% – 13%)
C only	4.1 (35% – 9%)	3.4 (35% – 7%)	4.6 (38% – 11%)
RCA			
All	3.0 (54% – 3%)	2.7 (53% – 1%)	3.2 (59% – 3%)
A only	6.2 (64% – 5%)	6.5 (62% – 5%)	6.8 (66% – 5%)
B only	4.9 (53% – 12%)	4.2 (52% – 7%)	5.5 (56% – 12%)
C only	5.0 (53% – 10%)	4.0 (52% – 7%)	5.4 (55% – 12%)

Table 6.8. *True and reconstructed migrations for the source–sink simulations*

	No advantage	Source advantage	Sink advantage
Real *A*→*B,C*	15.3	15.6	14.9
Reconstructed *A*→*B,C*	16.7	19.7	14.8
Reconstructed *B,C*→*A*	8.8	9.9	7.7
Total reconstructed	25.5	29.7	22.4

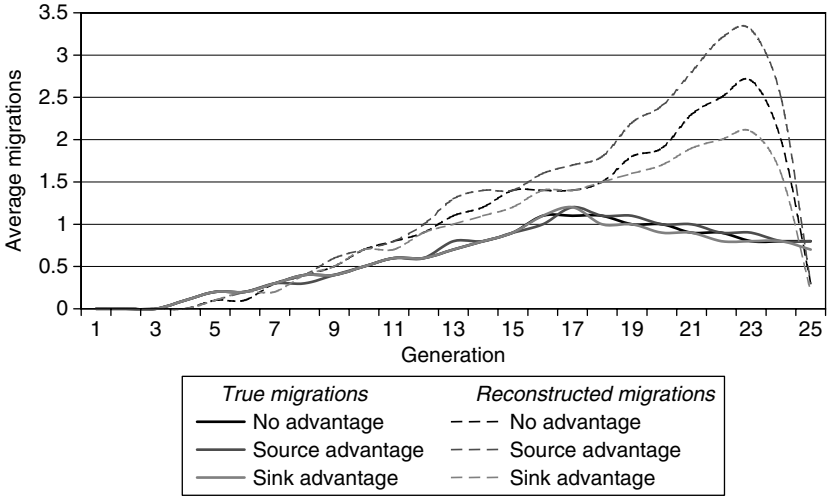


Figure 6.17. True and reconstructed migration profiles for the source–sink migration simulations.

count overestimates the number of migrations because the current generation on the destination continents frequently have their most recent fossil common ancestor on the source continent. This means that each current species on the destination continent adds one to the migration count, although it is clear that the migrant is, in essentially all cases, a species on the same continent that was not fossilised. The reconstruction manages to avoid this problem to a large degree by usually identifying a fossil as an ancestor. Although such an identification is in error under these circumstances, the net effect may well be to produce a more accurate phylogeny, because the chosen fossil is often closely related to the truly ancestral species, and thus the reconstructed migration pattern may therefore better reflect the true migration situation.

6.4 Unrestricted migration with advantage

The final migration scenario considered uses all four continents, each with its own initial species, and allows unrestricted migration between them. However, one continent is given the maximum possible selective advantage. Figure 6.18 shows the results for 25 generations, with a dashed horizontal

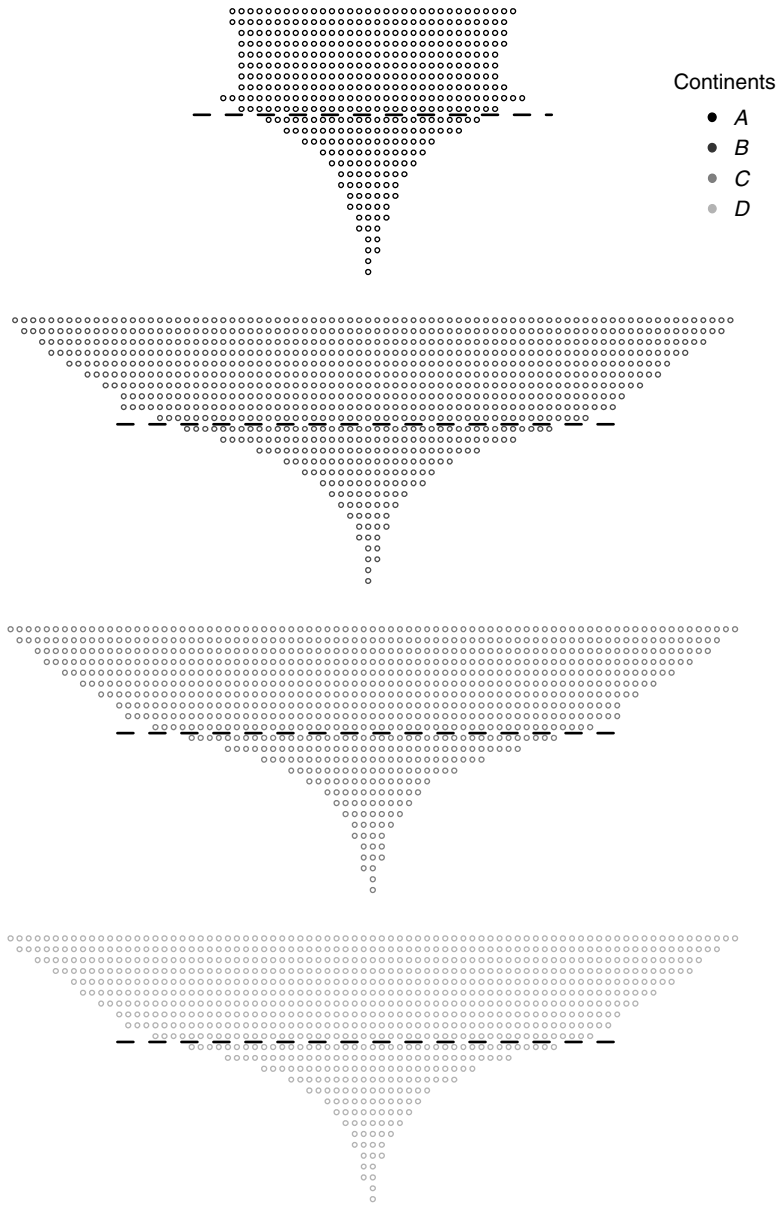


Figure 6.18. Average profiles for unrestricted migration over 25 generations. The dashed horizontal line on each continent indicates the fifteenth generation, about which time the differing selective advantage between continent *A* and the other continents is just beginning to have a significant effect.

line on each continent indicating the fifteenth generation. After 15 generations, there is never an overall common ancestor, and in more than 98% of cases there is no common ancestor for any individual continent. Fossil common ancestors are even rarer, but the reconstruction has a fairly even 'background noise' rate of between 2.9% and 6.2% of runs finding a reconstruction common ancestor, overall and for each continent. When the common ancestor exists, it is always right back at the start of the simulation, even in the reconstruction.

There is a very high percentage of fossils on extant lineages (72%), increasing to 89% for the reconstruction. The Wagner common ancestor finds an overall common ancestor at generation −4.1, i.e. approximately four generations before the start of the simulation. This is quite reasonable given that all four continents started with a single related species. However, the single-continent Wagner common ancestor estimates are generation 12 for continent *A* and 10 for the other three continents: quite a long way from the true value.

When all 25 generations are included, there is time for the selective advantage to take effect and species from continent *A* begin to dominate the evolution on other continents. In roughly 20% of cases, there is a common ancestor on continent *A* for each of the other continents. Still, in the other 80% of cases there is no common ancestor at all. This ratio is reduced when considering fossils only: down to about 5% because of sampling effects. The reconstruction is unable to pick up this effect at all, and still has quite a flat distribution, i.e. it is unable to determine the action of any selective advantage in the simulations. The Wagner overall common ancestor is at generation −10.7, certainly too far back in time. The percentage of fossil taxa connected to current taxa drops to 31% (61% according to the reconstruction) from 72% (89% according to the reconstruction) as a result of increased lineage-sorting in the longer run.

As the migrations begin to take effect, the advantage of the continent *A* taxa over taxa from the other continents shows up in the enhanced fitness of the migrants, and there is a corresponding diversity increase in these other continents. The higher species population then leads to more migration from these continents: approximately 46 migrations on average from each to any of the others, compared with an average only 18 migrations from continent *A* (see Table 6.9). The impact on the reconstruction results of the overall species population per continent, and thus the chance of finding representative fossils, is interesting. An overestimate is seen for migrations from continent *A*, but the number of migrations to the other continents is underestimated, especially migrations to continent *A*. Although not as obvious, this pattern is already evident in the results to 15 generations.

Table 6.9. *Average number of real and reconstructed migrations for the unrestricted migration simulation with selective advantage for species from continent* A

	Real	Reconstructed
From *A* to others	18.5	22.0
To *A* from others	45.9	16.5
To others from others	45.8	31.6

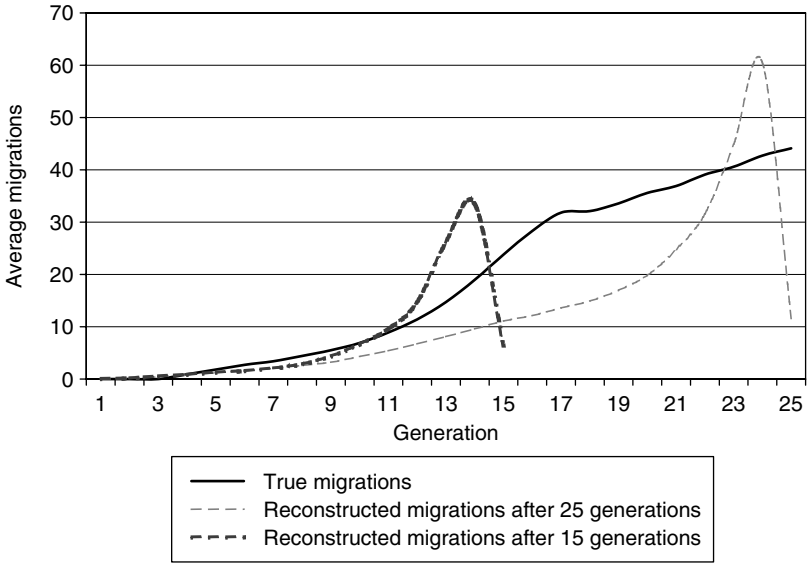


Figure 6.19. Migration profiles for unrestricted migration case.

The difference between these true and reconstructed migration profiles and those of the other cases studied in this chapter is also clear from Figure 6.19. The 'end effect' for the reconstruction, arising from migrations that end with current species, is seen in the peaks at the end of each of the reconstruction curves. After 15 generations, the reconstruction overestimates the total number of migrations slightly (approximately 8%), but after 25 generations there is an underestimate of more than one third of all migrations. The discrepancy occurs primarily in the middle period of the simulation, where the reconstructed migration curve stays well below the true curve.

The degree of monophyly in the identified clades is very low, less than 5% for 15 generations, but increases to 31% for 25 generations when the lineage-sorting has started to take effect. For current taxa, the degree of monophyly is less than 15% after the first 15 generations, but increases to 46% over the full 25. In both cases there is less polyphyly than paraphyly. Comparing with the Wagner reconstruction, there is only 2% monophyly after 15 generations, increasing to just over 17% after 25.

7 *Discussion*

The simulations presented in the previous two chapters have covered the generation and evolution of character diversity for both species and interbreeding groups, spread across up to four continents, and able to migrate in controlled ways. The simulation employs two specific techniques of phylogenetic reconstruction: one a method based on the morphological characters of current species and fossils, and the other a distance method based on the characters of the current species only. Particular emphasis was placed on the way in which the current species and fossils are employed by these reconstruction methods, and on the impact on their accuracy of both a sparse fossil record and the complex interplay between hereditary and non-hereditary effects on morphological characters.

Although the simulation is able to model evolution quite generally, in most cases constraints were placed on the evolution in order to more closely model important aspects of hominoid evolution. An overview of the results follows, with discussion of some specific implications, and finally future work and possible enhancements to the simulation are presented. In all cases where average results are discussed, the averages are over 1000 independent simulations.

7.1 Single-continent summary

The single-continent simulations presented covered species profiles with a recent minimum of species, subspecies profiles with a recent maximum of subspecies, as well as provided detailed analysis of the influence of fossilisation rates and non-hereditary characters. The main focus was on determining the time to the most recent common ancestor, the degree to which fossils lie on surviving lineages, and the accuracy of reconstructions. Detailed summaries of the results follow.

7.1.1 Species simulation results

In the absence of migration, three profiles with a recent reduction in diversity were studied. Specifically, these were a *vase* profile, where an approximately

linear increase in diversity is followed by a linear decrease; an *amphora* profile, where an initial increase in diversity is followed by a long 'neck' of reduced diversity; and a *mass extinction* profile, where after a period of increasing diversity, there is a sudden and drastic decrease. The fossil distributions produced by any single simulation were very similar for all three profiles, although each profile was capable of producing output that varied over a wide range, even when constraints were applied. Nevertheless, as subsequent analysis revealed, each of the different profiles produced phylogenies with its own distinguishing characteristics.

The most recent common ancestor was the furthest back for the vase profile, on average 13 generations back, compared with fewer than 8 generations back on average for the amphora profile. This was a result of the different degree of narrowness in the respective profiles for the more recent generations, and the corresponding effect on the number of surviving lineages. Related to this was the difference in character diversity among the current species for the different profiles, with the vase profile producing the most diverse current species and the amphora the least.

After generating each phylogeny, the focus moved to the reconstruction process, in particular the accuracy of the two methods employed when determining common ancestry. Problems in the reconstruction of the phylogeny arose largely from the presence of non-hereditary characters, and from the difficulty in estimating the character mutation rate. General ancestor–descendant species connection inaccuracies occurred frequently, affecting the recognition of potential ancestry and thus the identification of clades. In addition, the true number of fossils on extant lineages was always very low and was overestimated by about a factor of 2 by the reconstructions in nearly all cases. The reconstruction usually identified too recent a fossil as the most recent common ancestor, but, for the settings employed, on average this chosen fossil differed in time from the true fossil ancestor by only a few generations. The vast majority of clades identified were actually polyphyletic, indicating that the algorithm was overly inclusive. Clades containing the current species were identified very accurately overall, with more than 75% of reconstructed clades being truly monophyletic groups.

The Wagner reconstruction was quite accurate for determining the time of the most recent common ancestor. This is quite important, because the fossil reconstruction can only attempt to find the time of the most recent *fossil* common ancestor: generally much earlier in the simulation. However, the Wagner method was less successful than was the fossil reconstruction in determining the current species clades, highlighting the difficulty in identifying relationships between current species accurately without specific reference to fossil data.

7.1.2 Simulation results with lineage-merging

When interbreeding was included in the runs, the simulation units (taxa) corresponded to subspecies or interbreeding groups. In these cases, the increasing diversity over time was modelled by using either a *bowl* profile, where a constant species extinction rate of 25% results in a diversity that increases each generation, or a *logistic* profile, where the diversity increases until it reaches an equilibrium value. These profiles led to a common ancestor many simulation steps further back than was found for the profiles used in the species simulations, but it must be remembered that, for simulations with perhaps one or two hundred thousand years, rather than around one million years as for the species simulations. The degree of current species character diversity generated was quite similar for the two profiles.

The relationships between the time identified for the reconstructed common ancestor, the true fossil common ancestor and the true common ancestor are similar to those in the species simulation case, but the Wagner common ancestor results are very much poorer, greatly overestimating the time required to generate the current diversity for both profiles.

Compared with the species simulation, the runs with interbreeding showed a much higher percentage of fossils lying on extant lineages, as a result of there being more taxa overall, and a greater degree of interconnection. However, as was the case for the species simulation, this figure was again overestimated by the fossil reconstruction, this time by approximately 50%. The overall reconstruction accuracy was similar for the two kinds of simulation, but the clade identification accuracy was greatly reduced with interbreeding included, both when considering all taxa and when considering current taxa only.

7.1.3 Parameter sensitivity analysis

In order to understand these initial results better, it was necessary to study the sensitivity of the reconstruction methods to the two most important sets of parameters: those controlling the fossilisation rate and those controlling the nature of the non-hereditary characters. From this point on, all species simulations were restricted to either the vase or the amphora profile, and all simulations with interbreeding used the logistic profile.

Running simulations with the fossilisation rate set at a constant 100%, meaning that all taxa were available to the reconstruction, and with all characters hereditary and thus directly providing information about ancestry,

allowed the fundamental accuracy of the reconstruction and clade identification algorithms to be studied. The performance was very good, with the common ancestor generations completely correct, and the fossil connections

managed better than 83% monophyly in all cases; for the current species only clades, monophyly was as high as 98%. With interbreeding added, the identified clades in the fossil reconstruction only managed 63% true monophyly, but this increased to over 90% for current species clades. The only negative was the poor performance of the Wagner reconstruction, matching the highly successful fossil reconstruction in less than 30% of the runs with interbreeding.

The overall success of these runs implies that any errors and limitations in the reconstructions for the more realistic simulations are not intrinsic to the methods, but actually reflect the difficulties introduced by low fossilisation rates and non-hereditary characters.

The dependence of the results on the fossilisation rate was studied by considering runs with rates of 3%, 10% and 30%, as well as runs where the rate varied from 5% for the early generations up to 15% for the late generations. Interestingly, the percentage of fossils on current lineages was largely independent of the fossilisation rate, staying constant at 17% for the

lower the fossilisation rate, the greater the overestimation of this figure in the reconstruction: by as much as a factor of 3 for the amphora profile with a constant 3% fossilisation. This was part of a general degradation of the connection accuracy in the reconstruction as the fossilisation rate decreased, because the absence of fossils decreased the degree of discrimination that the reconstruction was able to apply to fossil species. As expected, the true fossil ancestor was earlier in the simulation when the fossilisation rate was lower, but, surprisingly, the reconstruction failed to capture this fact, and instead wrongly labelled a fossil as common ancestor from roughly the same generation in all cases. For the species simulation, the current species clade identification remained quite successful, although it did degrade as fossilisation decreased. However, for the logistic profile with interbreeding, the clade identification performance became very poor at the lower fossilisation rates, only managing true monophyly in 30% of the identified current species clades for the lowest fossilisation rate of 3%.

Of equal importance is the dependence of the results on the nature of any non-hereditary characters, and this was studied by considering runs with 25% and 50% of characters non-hereditary, and with a likelihood of change of 20% and 50%. Unlike the case above, where the Wagner common-ancestor and reconstruction common-ancestor estimates were essentially independent

of the fossilisation rate, for these runs the accuracy of both these estimates decreased with the increasing role of non-hereditary characters, whereas the true fossil common-ancestor generation was unaffected, as was the case for all fossil measures. An important way in which the results were affected by this change was the introduction of greater error in the estimate of the character mutation rate for both reconstruction methods. Because of this, the increased presence of non-hereditary characters reduced the connection accuracy of the reconstruction, and thus also seriously affected the identification of clades. Consistent with the earlier cases, the most serious degradation was for the logistic profile with interbreeding.

7.2 Migration summary

The introduction of migration greatly increases the range of situations that can be simulated. Five particular scenarios were studied in Chapter 6: one where the migrations were restricted in time; another where the migrations were restricted in direction; one where the migrations were constrained to follow the pattern seen in hominoid evolution; and two where the migrations were unrestricted. Because the results in Chapter 5 had already covered the issues of non-hereditary characters and fossilisation rates in detail, the main focus of the migration results was the recovery of the different migration patterns and the role of selective advantage.

An immediate impact of the introduction of evolution on distinct continents was greater current species diversity, because the associated isolation made independent lineages more likely to persist. Given time, reciprocal monophyly usually developed, but, of course, this could be undone by just a single migration. For these reasons, the ability of the reconstructions to determine migration history was of primary importance. However, doing so, especially where the non-hereditary characters are similar between the existing and migrant species, is a very difficult task.

For two continents, labelled *A* and *B*, constrained to evolve according to an amphora diversity profile, with just continent *A* populated initially and a 5% chance of migration between them, the fossil reconstruction had some difficulty in determining the location of the common ancestor for each continent. For example, on continent *B*, the common ancestor of its current species was from continent *A* for only 13% of the runs and from continent *B* for 79% of the runs. (The remaining runs resulted in no current population on this continent.) This distribution was slightly distorted by the fossilisation rate, in this case varying linearly from 10% to 30%, and so the fossil common

ancestor was on continent *A* for 17% of the runs and on continent *B* for 69% of the runs with 14% of the runs producing no fossil common ancestor for continent *B* at all. According to the reconstruction, the fossil common ancestor was on continent *A* for 29% of the runs and on continent *B* for only 56% of the runs, so instead of finding the true ratio where a continent-*A* location for the continent-*B* common ancestor was four times less likely than a continent-*B* location, the reconstruction found a ratio of less than 2. The continent-*A* common-ancestor location result also illustrates how fossilisation can distort the migration picture, because the true continent-*A* ancestor was on continent *A* for 92% of the runs and on continent *B* for only 2% of the runs, whereas the fossil common ancestor for continent-*A* species was on continent *A* for only 79% of the runs and on continent *B* for 9% of the runs. This similar, but significantly different, picture is all that the fossil reconstruction can hope to uncover.

The simulations in Section 6.2 modelled hominoid migrations by simulating a source continent (continent *A*, Africa), and a period of two-way migrations via an intermediate continent (continent *B*, Europe) to a third continent (continent *C*, Asia). Species on the intermediate continent were made extinct at the end of the period of migration.

The results of a single run were shown first, revealing a number of interesting features. Fossils were evenly distributed across the three continents, but there was an anomalously high current species character diversity on continent *C*, and the most recent common ancestor was on continent *B*, i.e. the continent with no surviving species. The overall reconstruction accuracy was only around 20%, although all current species were correctly reconstructed. Most importantly, however, of the 15 true migrations, *none* was respresented in the fossil record, i.e. no fossil species was on a lineage that included a fossil on another continent! Subsequently, the reconstruction was unable to identify either of the two periods of greater migratory activity.

With the average results the current species character diversity on continents *A* and *C* matched, as would be expected, and there was a recent common ancestor for each of these continents, slightly more recent on average on continent *C*, i.e. the Asian common ancestor. The location of the single-continent common ancestor for continent *A* was always also on continent *A*, and for continent *C* it was also on continent *C* approximately 95% of the time. However, the overall common ancestor was back before generation 5, i.e. prior to the start of the migrations, and therefore nearly always on continent *A* (93% of the time) although one run in the 1000 had the most recent common ancestor of all current species on continent *C*!

There was an overall fossil common ancestor for only 22% of runs, although an individual-continent fossil common ancestor was present in more

than 80% of runs, for both continents. There was a significant possibility of the continent *C* fossil common ancestor occurring on continent *A* or continent *B*, but only a small chance of the continent-*A* fossil common ancestor being a fossil on continent *B*, and almost no chance of its being a fossil on continent *C*. As has been found consistently in the species simulations, only a small percentage of fossils were found to lie on surviving lineages (20% in this case) and the reconstruction approximately doubles this percentage (39% in this case). Of particular interest is the fact that only 6% of fossils on continent *B* were found to lie on surviving lineages, despite the fact that more than a quarter of all fossils are on that continent. In contrast, although only 23% of fossils were on continent *C*, 29% of these were on surviving lineages. This illustrates the difficulty introduced by the period of extinction on continent *B*, which, given the fundamental importance of this continent in the overall development of the current species situation, is clearly a serious problem.

When migrations were restricted in various ways, the incompleteness of the fossil record made it is very difficult for the reconstruction to capture the nature of the restriction. This was especially clear in the simulations where there was only a small window in which migration was possible, a situation that is quite likely to be a recurring pattern in human history. Several simulations with interbreeding and migration between continents *A* and *B* were run, again with continent *B* initially unpopulated, with four different migration windows, each of size six generations. The reconstruction was unable to recognise the restrictions that had been placed on the migrations, essentially smoothing the migrations out over the period from a little before the start of the window until the final generation of the simulation. In fact, for the case where migrations were only possible between generations 5 and 10, the reconstruction indicated that migrations had occurred at an almost constant (and maximum) rate from generation 10 right through to around generation 23. The pairwise character differences amongst the final population, both overall and on continent *A* only, were roughly constant in all cases. Considering continent *B* only, the difference increased substantially with the recency of the migration window: from 7.3 characters per species for the earliest window up to 9.5 characters per species for the latest window.

In order to study more closely the difficulty the reconstruction had in recovering time-limit constraints on migration windows, the average results for runs with a migration rate of 5% between generations 10 and 15 were compared with those for a migration rate of 2% between generations 5 and 20. These two situations produced a similar degree of total migration, and the true migration profile for the second case was essentially the same as the profile determined by the reconstruction for the first case. The primary distinguishing feature between these two cases was that the percentage of runs where the

common ancestor for continent *B* was on continent *B* was significantly lower for the second case, for the true common ancestor, the fossil common ancestor and the reconstructed common ancestor.

When the restriction was one of direction rather than time, as is the case for source–sink migrations, the nature of the results was again greatly changed. In particular, the time to the common ancestor increased when the *source*-continent taxa had a selective advantage over the *sink*-continent taxa, relative to the situation when there was no advantage. Conversely, the time to the common ancestor was reduced when the sink taxa had a selective advantage over the source taxa. This was true for the fossil ancestor and the reconstruction ancestor estimate as well, except for the common ancestor of the source-continent taxa only, where the generation of the common ancestor varied very little. Similarly, the degree of pairwise character difference between current taxa was more or less constant for continent *A* in all three cases, but increased on the sink continents when the source continent had the selective advantage, because the migrant taxa were more likely to become established and thus the introduced diversity was more likely to persist. The character differences between current taxa decreased when the sink continent had the advantage. The reconstruction in each source–sink scenario introduced many spurious back-migrations, with the source advantage case leading to the greatest overestimate, owing primarily to the larger number of taxa.

The barrier-migration and source–sink simulations report very high numbers of migrations in the true fossil record: higher than both the true number of migrations and the number of reconstructed migrations. This results from a combination of the relatively large number of current species and the fact that, when the immediate fossil ancestor of a group of current species is on another continent, each current species on the destination continent adds one to the migration count, although it is clear that the migrant is, in essentially all cases, a species on the same continent that was not fossilised. In the fossil reconstruction, the more frequent identification of a fossil as ancestor means that this problem is avoided to a large degree, and although such an identification is frequently in error the net effect is to produce a more accurate phylogeny, because the chosen fossil is often closely related to the truly ancestral species and thus the reconstructed migration pattern may therefore better reflect the true migration situation.

The Wagner reconstruction was unable to cope with the migration and interbreeding runs at all, producing estimates for the time to the common ancestor that were far wrong, both overall and for any single continent.

The final situation simulated involved unrestricted migration across four continents, each with its own initial species, but with the maximum possible selective advantage for continent *A* over the other three continents. With

respect to the recovery of migrations by the reconstruction, rather than the pattern of overestimation seen in nearly all the simulations discussed above, the reconstruction in this situation was unable to recover enough migrations, because they were insufficiently represented in the sparse fossil record. Thus interpreting migration data as indicated by the reconstruction requires, to some extent, independent prior knowledge of the degree of migration. In a sense, it seems you have to know the answer to get the answer!

7.3 Implications

Some recurring themes are apparent in the above discussion. With respect to the fossil reconstruction, there is the persistent overestimation of the number of fossils that lie on extant lineages, the influence of the species fossilisation rate on the general accuracy of the reconstruction, and the difficulties arising from the presence of non-hereditary characters. The non-hereditary characters are even more problematic for the Wagner reconstruction, especially when there is merging of lineages through interbreeding. All of these problems reflect genuine difficulties in phylogenetic reconstruction. Species fossilisation rates are sure to be low in general (Tavaré *et al*., 2002), and thus limit the availability of fossil character information. Non-hereditary influences on the available fossil characters are a significant problem (Oxnard, 2000), especially when some species may be known from only very few fossils. In addition, given the difficulties with the concept of *species* discussed in Chapter 2, methodological and practical complications due to the definition and merging of lineages cannot be ignored.

In the hominoid situation of many prior species and relatively few current species, the problem of determining exactly to which lineage a particular fossil belongs, or indeed whether a particular fossil, although similar to a current form, is or is not a direct ancestor, is of substantial interest. Over the past 10 million years or so of hominoid evolution, this issue is particularly significant, both because of the human–chimpanzee–gorilla (near) trichotomy and because of the special interest in finding the *first human*. The simulation results illustrate how large differences can result from relatively small errors in this regard.

As discussed in Section 2.2, the model of human evolution that supposes a straightforward linear progression through various forms has been seriously compromised by a number of discoveries over the past decade. Although the additions of *Paranthropus* (Wood and Collard, 1999), *Ardipithecus* (Haile-Selassie, 2001; White *et al*., 1994) and maybe even *Orrorin*

(Senut *et al*., 2001) could still be accommodated within a relatively simple branching model, *Kenyanthropus* (Leakey *et al*., 2001) and most significantly *Sahelanthropus* (Brunet *et al*., 2002) seem to defy any such solution. These two genera consist of a combination of features found other genera, in some cases genera millions of years more recent. For example, the canines and brow ridge of *Sahelanthropus tchadensis* appear most closely related to those found in species of the genus *Homo*. Wood (2002) speaks of the *untidy* model of human origins, where characters are 'mixed and matched' in successive adaptive radiations. This is certainly consistent with the situation implied by the simulations, of many species with similar characters, but mostly on extinct lineages.

Other important issues are those related to the effects of migration and selective advantage. Migrations in particular are a very serious complication, greatly influencing the results yet very difficult to unravel from the fossil record. The simulations have shown how hard it can be to determine the nature of any historical restrictions on species migration, and even possible difficulties in determining whether the fossil record suggests an overestimate or an underestimate of the true degree of migration. With early hominid species spread across Africa, and human species distributed globally, the degree of hominid mobility is clear. Even earlier, the fact that hominoid species are found entirely in Africa during the early Miocene, but then mainly in Asia and Europe during the middle to late Miocene, shows the importance of understanding the role of migration if common ancestry is to be accurately reconstructed. Furthermore, independent and mobile groups of humans are certainly potentially interbreeding, even though differing environments and prior reproductive isolation may have led to a degree of difference in their characters, both hereditary and non-hereditary. The question is more one of when such changes are sufficient to block interbreeding. The simulations show the extent to which even quite a small degree of interbreeding greatly complicates the reconstruction process.

In short, although various reconstruction methods may be used to justify different phylogenies, the precise nature and extent of the impact of the above complications indicates the problems associated with assuming that any particular reconstruction actually reflects the true situation. The difficulty in determining the relationships between fossils, and in particular the tendency to place fossils on current lineages far too frequently and to connect current forms to fossils that are too recent, is a primary source of error in hypothesising ancestor–descendant relationships. Even in the absence of interbreeding, the complications in identifying the nature of characters, and similarity in characters between different species, imply that great caution must be exercised when identifying common ancestors and associating a

time with them. In general, it seems that quite subtle features in the data are of central importance, and all available alternative data sources need to be exploited in order to try and identify the broad framework underlying the evolution, such as possible migratory history and diversity patterns, with standard methods only then being applied with these limitations in mind.

Although I have attempted to cover a wide range of important situations via the simulations presented in the two previous chapters, these are in fact only a very small subset of the possible situations the simulation is capable of modelling. For example, another interesting case to focus on for the study of hominid origins is the amphora profile, but with greater or lesser *constriction* in the neck region of the profile. This would be possible by a simple adjustment to the extinction function, and would enable focussing on either the phylogeny for all extant hominoids, or only the African ape clade. Alternative profiles would enable modelling of primate species generally; indeed, with little or no modification, the simulation may be adapted to any number of different species and evolutionary scenarios (Oxnard and Wessen, 2001).

7.4 Future work

An obvious area for refinement is in the phylogenetic reconstructions. For example, it may be possible to include fossil species in the Wagner reconstruction and to try more sophisticated reconstruction techniques. In addition, more precise measures of the similarity of the true and reconstructed phylogenies are possible, and will provide more information than the current comparison of successful connections and clades. Furthermore, a modified and more inclusive clade-identification algorithm may be appropriate for the simulations with interbreeding.

Seeing that both the species and interbreeding groups simulations are actually simulating species evolution, the ability to *collapse* the simulated phylogeny from a run with interbreeding into its corresponding species tree, by superimposing merged lineages, would be very useful for bringing together results from the two different kinds of simulation.

Finally, more sophisticated modelling of the environment and associated functional adaptation, both hereditary and non-hereditary, could greatly increase the applicability of the simulation results to real world situations. One particular way that seems promising involves treating sets of characters as composites, with the most significant members of the set determining the group identity and the least significant members allowing change in detail without changing the overall interpretation of the function of the whole set.

Here, significance is essentially a measure of the degree of resistance to evolutionary change. Both the mutation model and the non-hereditary character model would have to change to take into account the different levels of significance of different characters. Perhaps the most interesting possibility this more realistic model would provide is the ability to study approaches for attempting to differentiate between hereditary and non-hereditary characters automatically in the reconstruction.

This first part of this book has been mostly concerned with the construction of a relatively general model of species and subspecies evolution, the identification of important parameters, and the analysis of some preliminary results. Future applications will involve focussing on more specific situations, and then applying more sophisticated statistical methods to enable the study of the physical and evolutionary issues of relevance in greater depth.

Part II
Simulating genealogies

8 *Overview*

In the first part of this book, simulations designed to generate a tree of related but separate species were studied. The primary concern was the reconstruction of phylogenies from the simulated fossils, and how the accuracy of the reconstructions was affected by factors such as the rate of fossilisation, non-hereditary adaptation and migrations. The discrete species generations represented time steps ranging from a few hundred thousand years to one million years, and the simulation resolution was, at finest, subspecies or large interbreeding groups. In this second part, the simulation units are individuals and the time steps single, non-overlapping generations. Thus genealogies, rather than phylogenies, are simulated. In order to allow for a fine and detailed analysis of the role of various parameters of interest, e.g. sex ratio and breeding patterns, varying and structured populations, these simulations must involve much larger populations and many more generations than in the species case.

Neutral models permit the separation of demographic effects from mutation (Kimura, 1968); because demographic effects, such as those mentioned above, are the primary focus of the simulations, introduction of the genetics models is delayed until Chapter 12 and the reconstruction of genealogies is assumed to be perfect. Incorporating selection substantially increases the complexity of any analysis, precisely because such a separation is not possible (Hudson, 1990; Tavaré *et al*., 1997). Nevertheless, selective advantage is included in the genetics model to a small degree.

The use of mtDNA and Y chromosomes for tracing human genetic ancestry was discussed in Section 1.2. Mitochondrial DNA and the Y chromosome are inherited from a parent of a particular sex only, and so, if we look back in time, any individual has only a single ancestor in the previous generation. In fact, in such a population, termed *haploid*, any current individual has only a single ancestor in any previous generation, and as we move back in time there will be a gradual random merging and associated loss of lineages. The rate at which this occurs is the described by *coalescent theory* (Kingman, 1982a,b), an outline of which follows.

8.1 Coalescent theory

Consider any two individuals in a haploid population with a constant over time *effective* population of N, i.e. an ideal population of N breeding individuals with genetic properties approximating those of a larger, real population. (The relationship between the effective population and the total, or *census*, population will be discussed in more detail below.) Assuming an equal probability for individuals from the previous generation to be the parent of any current individual (when N is large, this implies a Poisson distribution in number of offspring (Avise, 2000; Hudson, 1990)) the probability that the two chosen individuals share a parent is simply

$$P_1 = \frac{1}{N}.$$

Going one step further, the probability that their most recent common ancestor is a grandparent is given by

$$P_2 = \frac{1}{N}\left(1 - \frac{1}{N}\right),$$

i.e. the probability that they do not share a parent multiplied by the probability that their distinct parents do. In general, the probability that the lineages remain distinct for exactly t generations is

$$P_{t+1} = \frac{1}{N}\left(1 - \frac{1}{N}\right)^t, \tag{8.1}$$

which in the limit of large N becomes

$$P_{t+1} \approx \frac{1}{N}\mathrm{e}^{-t/N}, \tag{8.2}$$

showing that the probability is approximately exponentially distributed with mean N. Therefore the expected time for two arbitrary lineages in a haploid population of constant size N to *coalesce* is N generations.

The problem of determining the time to the most recent common ancestor of an entire generation is more complicated: the presentation below follows that of Hudson (1990). Again assuming a constant effective population of size N, where the parent of any individual is equally likely to be any individual from the previous generation, the probability for two individuals is as calculated above. To extend this result, note that if k chosen individuals have distinct

parents with probability $P(k)$, then the probability that $k+1$ individuals have distinct parents is given by

$$P(k+1) = P(k)\left(1 - \frac{k}{N}\right).$$

In general, the probability that m sampled individuals have m distinct parents is

$$P(m) = \prod_{k=1}^{m-1}\left(1 - \frac{k}{N}\right) \approx 1 - \frac{1}{N}\binom{m}{2}, \tag{8.3}$$

where the approximation is again for the large N limit.

This result applies as much to any prior generation as to the current one, and so the probability that m sampled lineages remain distinct for exactly t preceding generations (i.e. the first coalescence occurs $t+1$ generations back) is given by

$$P(m)^t(1 - P(m)) \approx \frac{1}{N}\binom{m}{2}\mathrm{e}^{-\frac{1}{N}\binom{m}{2}t}. \tag{8.4}$$

This is an exponential distribution with mean $N/\binom{m}{2}$.

For $N \gg m$, there is very little chance of two lineages coalescing, and thus a common ancestor occurring, in a single generation. The expected time over which j lineages persist is given by

$$E(T(j)) = \frac{N}{\binom{j}{2}} = \frac{2N}{j(j-1)}, \tag{8.5}$$

and thus the expected number of generations to the overall common ancestor, i.e. when the last two distinct lineages merge, is

$$T_{\mathrm{MRCA}} = \sum_{j=2}^{m} E(T(j)) = N\sum_{j=2}^{m}\frac{1}{\binom{j}{2}} = 2N\left(1 - \frac{1}{m}\right). \tag{8.6}$$

When $N \gg m \gg 1$, this results in a time of approximately $2N$ generations to the common ancestor, with variance

$$\mathrm{Var}(T_{\mathrm{MRCA}}) = N^2\sum_{j=2}^{m}\frac{1}{\binom{j}{2}^2} \approx 1.16N^2. \tag{8.7}$$

Note that $E(T(2)) = N$ (as can be seen from Equation 8.5 above, or from Equation 8.2 of the earlier calculation) and thus more than half the total

coalescent time is the time taken for the final two lineages to merge. A side-effect of this is that, although the coalescent tree carries all the genetic history of the current population, it has relatively little information about the demography of the population for a substantial portion of the time back to the common ancestor. As will be discussed in the following two sections, demographic information from a general coalescent tree is only available indirectly, e.g. by examining the branching patterns, possibly of various trees, and making deductions as to the nature of the historical population in terms of size, sex ratio, mating patterns, fertility, population structure and migration, the effects of selection, etc. A question of great importance then concerns the degree to which this situation persists for more realistic models, and thus how much of a limitation it is in the practical application of coalescent models.

Equation 8.7 shows that nearly all the random variation in the genealogy is in this two-branch period (Harding, 1996; Tavaré *et al*., 1997) and that the standard deviation remains large, of order N, even as the sample size is increased. A further interesting result is that the probability that a sample of size m leads to the most recent common ancestor of the whole population is given by $(m-1)/(m+1)$, and so the overall common ancestor is very likely to be found even when only a small sample is used (Saunders *et al*., 1984).

If an additional factor, $0 < \sigma^2 < \infty$, describing the variance in number of offspring, is included, the right-hand sides of Equations 8.5 and 8.6 need to be divided by σ^2. (The $\sigma^2 = 1$ model, as presented above, is the well-known Wright–Fisher model.) This factor implies that for a larger variance in number of offspring the time back to the most recent common ancestor is shorter, and vice versa (Kingman, 1982a; Tavaré *et al*., 1997).

The situation for autosomal genes is similar to the above, because any particular gene is also inherited from a single parent, although at each generation there are two possible choices. This corresponds to a *diploid* population and, provided there is no recombination, simply requires the substitution $N \to 2N$ in the above calculation. However, in this case the effective population includes both sexes, so the net effect is a fourfold increase in the effective population and thus in the time to coalescence (Avise, 2000; Hudson, 1990; Palumbi *et al*., 2001).

Figure 8.1 shows a simple genealogy for an approximately constant population and highlights the transmission of mtDNA along matrilines, the transmission of the Y chromosome along patrilines and the arbitrary sex transmission of a sample gene on an autosome.

A number of assumptions that are unlikely to be met in practice underlie the above results. In particular, the model implies independence of lineages, absence of selection, no recombination, random choice of mates, and constant population. In reality, lineages are jointly affected by many external events

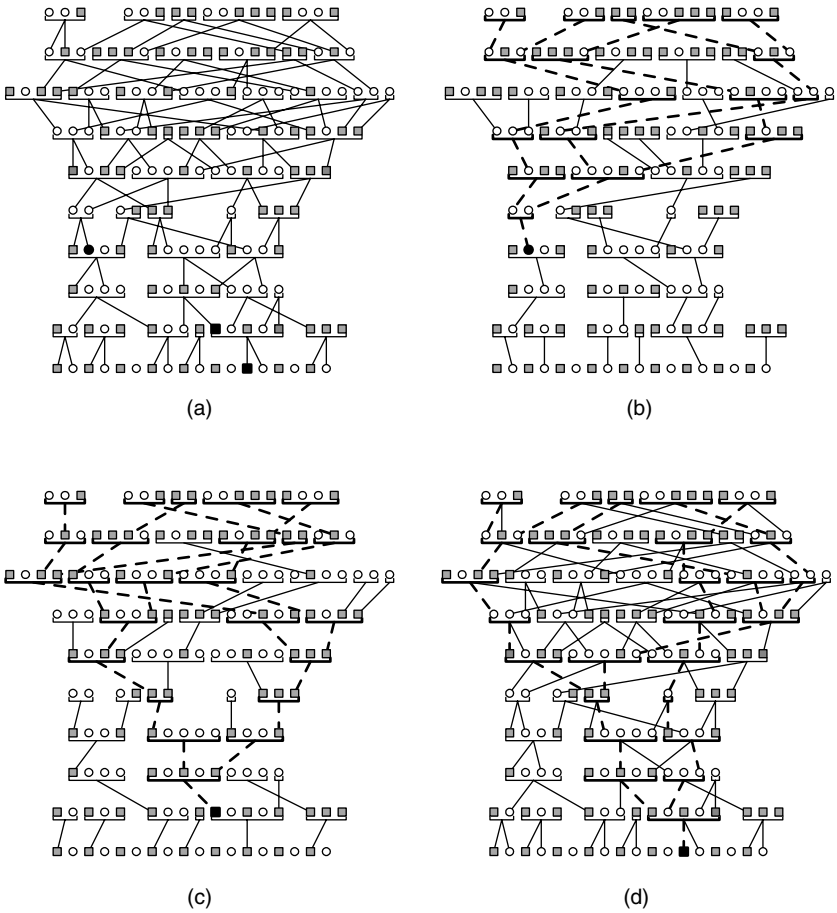


Figure 8.1. A simple genealogy, showing: (a) all relationships, (b) matrilines and most recent maternal common ancestor, (c) patrilines and most recent paternal common ancestor, (d) most recent common ancestor for an arbitrary autosomal gene. Males are indicated by shaded squares, females by circles, and the horizontal brackets delimit sets of siblings. The lineages highlighted by using a thicker dashed line are those surviving to the current generation, and the most recent common ancestor of the relevant type is shown coloured black in each figure. (All three most recent common ancestors are shaded black in (a).)

and are therefore not independent; recombination turns trees into graphs, which obviously has significant topological implications (Hudson, 1990); and selection changes the parent-choosing probability (from the perspective of looking back in time). Lineages further vary because of different

sex ratios, different breeding patterns and degrees of reproductive success, different dispersal patterns between populations and between sexes, and non-random mating, in particular due to population isolation (Avise, 2000). A more realistic model must consider these effects and many related complications.

Selection introduces substantial difficulties because it invalidates the basic assumption of a random choice of mates. Furthermore, different types of selection can have completely opposite effects on the genealogy. For example, balancing selection acts to maintain distinct alleles and thus distorts the genealogy in a way that increases both the mean and the variance of the time to the most recent common ancestor. This deepening of the genealogy is very similar to that resulting from strong population subdivision. Conversely, a selective sweep rapidly fixes a favourable allele, thus removing any existing polymorphism. This reduces the time to coalescence and produces a genealogy that is similar to one resulting from growth from a small population (Nordborg, 2001*a*). Coupled with these theoretical problems is the fundamental problem that it is not always apparent whether or not selection is acting at a locus, and if so, in what way.

The use of effective population rather than total population allows some of the demographic complications to be overcome. Different sex ratios and breeding patterns are essentially variations in the effective population size, or, equivalently, a variation in the time scale relevant to the above calculations, and thus act to change the rate of coalescence without fundamentally altering the basic results. When the number of breeding males in a population, N_m, equals the number of breeding females, N_f, the autosomal effective population is $N_{au} = N_m + N_f$. However, when this is not the case, the general expression for the autosomal effective population

$$N_{au} = \frac{4N_m N_f}{N_m + N_f} \tag{8.8}$$

must be used, and this has some interesting properties. For example, in a population where, on average, each breeding group consists of a single male plus three female mates, $N_f = 3N_m$, and substituting back into Equation 8.8, $N_{au} = 3N_m = N_f$, which is only 30% of the total number of breeding individuals, $N_m + 3N_f$. Thus, genetically, the breeding population appears far smaller than is actually the case.

Of greater significance is population structure, which has a direct impact on the shape of the coalescent tree because it acts as a barrier to the coalescence of particular lineages. Similarly, a fluctuating effective population changes the shape of the coalescent tree, because the rate of coalescence will depend on the precise nature of this time dependence. The census population may

be many times the size of the effective population (Relethford, 1998) or as little as twice as large (Sherry *et al.*, 1998), and this ratio may have been lower prior to the evolution of the long post-reproductive life span in humans (Gage, 1998).

8.2 The historical human population

Some models of modern human origins need a total population of hundreds of thousands to millions for the past one million years, whereas others only require the number of breeding individuals prior to the expansion (i.e. origin) of modern humans to be in the thousands. Human populations certainly have a history of expansion and contraction, and much effort has been spent trying to distinguish between possible population profiles. In particular, there is the hourglass theory that supposes a population of hundreds of thousands to millions prior to a Pleistocene contraction, and the long-neck hypothesis that supposes a persistent small population, of order 10 000, over the past million years or so (Harpending *et al.*, 1993). The absence of deep allelic divergence in nuclear genes in humans, relative to other hominoids, seems to support the long-neck hypothesis (Sherry *et al.*, 1998) but may be due simply to greater migration in human populations (Relethford, 1998). Rogers and Jorde (1995) review the genetic evidence and claim that, between 33 000 and 150 000 years ago, the human population expanded from 10 000 breeding individuals to a size of at least 300 000, and that tens of thousands of years prior to the expansion, small individual populations had already separated. An important point is that the effective population of a fluctuating population is much closer to the minimum than to the average (Avise, 2000; Harpending *et al.*, 1993), so the low effective population size may reflect this effect rather than low census size (Relethford, 1998).

Another approach can be found in the paper by Takahata and Satta (1997), who studied nuclear gene sequences of humans, chimpanzees, gorillas, Old World monkeys and New World monkeys, and determined an N of the order of 100 000 during the Pliocene and late Miocene, increasing to 1 million for the Oligocene and earlier. However, for the past few hundred thousand years their results indicate a lower population, in the range from 6000 to 17 000.

Extending the calculation in Section 8.1 to allow for variable population size shows that coalescent events are concentrated in the period prior to an expansion (Harpending *et al.*, 1998). In the particular case of exponential growth from a small population, the coalescent tree is stretched at the leaves and compressed near the root and therefore, relative to the constant population

tree, most coalescences occur early in the history of the population (Tavaré *et al*., 1997). This situation leaves the same signature in the coalescent tree as a selective sweep (discussed above) and leads to a reduction in the estimate of the size of the ancestral population. On the other hand, a population contraction leads to a loss of genetic diversity and in general to a much shallower tree. Nevertheless, some loci still carry variants retained from before the contraction, and thus the coalescent tree at these loci will still indicate the full depth. These factors underline why the assumption of neutrality is so important: if the locus is not neutral, any deductions regarding the effective population may be totally inaccurate owing to the action of selection. The use of data from several loci is one way to cross-check values (Relethford, 1998). Tavaré *et al*. (1997) warn that little practical information is known about the random processes governing the rate of birth in a population, and therefore that caution is always necessary when applying coalescent theory to realistic, variable populations.

In general, population growth has less impact on the time to the most recent common ancestor than prior population size and structure, and the precise nature of these in human history are the subject of much debate (Avise, 2000; Cann, 2001). In addition, in a geographically structured environment, different types of migration will have different effects on the genetic structure of the populations involved. For example, there may be voluntary movement to find mates or resources, sometimes hostile, or enforced movements such as slavery or wife-stealing, and the likelihood of migration within a population may depend on economic or social class. In recent history, written records and surnames provide information on migration in human populations, the latter for male migrations only. Data for human populations show generally small migration distances, especially in pre-modern European populations, but Indian and some hunter–gatherer populations, for example, show more mobility (Wijsman and Cavalli-Sforza, 1984). For populations in earlier times, archaeology and the evolution of language can provide information on the relatedness and migrations of populations, but for prehistoric populations genetic evidence must be relied on (Cann, 2001; Lewin, 1993). Although large-scale population movements are relatively rare, they will have a greater impact on the genetic structure, and given a specific hypothesis about the genetic origin of a population, gene frequencies can be used to estimate the degree of migration.

On balance, the genetic data seem to suggest an African expansion predating those of other major regional groups by tens of thousands of years, indicating the presence of a large African population for perhaps 100 000 years. As for selection, discussed above, different kinds of population structure can have opposite effects on the coalescent tree and related population estimates.

For example, isolated subpopulations extend the life of ancient lineages and thus increase the time to the coalescent, whereas a source–sink migration structure (see Section 6.3.2) can have the opposite effect, reducing the time to the coalescent for comparably sized populations (Avise, 2000; Tavaré *et al*., 1997).

Differences in the reproductive output or migration rate between the sexes will also be reflected in the geographic patterns and genetic diversity. Research by Seielstad *et al*. (1998) found that Y-chromosome variants are more localised than are mtDNA and autosomal variants, implying a greater migration rate in females (e.g. due to *patrilocality*, the custom in marriage where the wife goes to live in the husband's community). Other factors may also contribute to these differences. Polygyny can produce similar patterns (mating-pattern issues are discussed in more detail in Section 8.3 below); a higher rate of male mortality may also be a factor, as could be selective action on the Y chromosome. This underlines the difficulty in separating these effects based on genetic evidence alone.

In addition to the modelling difficulties described above, there is the significant difficulty arising from the large variance inherent in the coalescent analysis (Tavaré *et al*., 1997). As discussed above, the nature and statistical distribution of possible coalescent trees depends very sensitively on poorly known details of population structure and migration, and thus the tree determined from any single locus is not necessarily representative (Harding, 1996; Harpending *et al*., 1998; Satta *et al*., 2000). Furthermore, the human genome contains up to 40 000 genes (International Human Genome Sequencing Consortium, 2001) and each genomic region may have a different history of mutation, selection and recombination and thus a different effective population (Nordborg, 2001a), although it is important to remember that selection tends to act on individual genes whereas demographic effects are generally reflected more widely across the genome (Barbujani and Bertorelle, 2001).

8.3 Human mating patterns and fertility

While there remains substantial variation in human mating patterns and fertility rates, the current situation is greatly influenced by issues related to overall population size and growth rate. The total human population has increased from 1 billion in 1800 and 1.5 billion in 1950 to a massive 5.5 billion today. Lancaster (1997) describes how, parallel to this, the total fertility rate (TFR) has dropped to near or below replacement levels in all developed countries and is rapidly declining in developing countries: from 6.1 births per woman in the mid-1960s down to 3.8 in 1990. Various aspects of a developed lifestyle

also drastically alter the basic economic considerations historically associated with fertility in human populations.

In order to understand the mating patterns and fertility rates relevant to the study of ancient human populations, it is natural to study those groups whose lifestyle seems to correspond most closely with the supposed lifestyle of our ancestors. Despite the problems associated with using modern-day hunter–gatherer populations as analogues of the human population of the Pleistocene (O'Connell, 1999), they still stand out as the most reasonable candidates. With respect to fertility rates, there is substantial variation in modern hunter–gatherer populations, ranging from the !Kung of southern Africa with a TFR of just over 4.5 to the Aché of Paraguay with a TFR of more than 8 (Kaplan, 1994; Lancaster, 1997). Kaplan (1994) provides a model of fertility, modelling the relationship between actual fertility and child survival, and finds an optimal fertility rate, i.e. a fertility rate that maximises fitness, of approximately 5.77 births per woman.

There are limited historical records that can be used for fertility-rate calculations, but one case of interest comes from data on the British aristocracy covering more than 1000 years. These data show that, within this group, the fertility rate has ranged from 2.3 before 1500 to a maximum of 2.8 at the start of the seventeenth century (during the reign of Elizabeth I) and then steadily decreased to 1.5 by 1875 (Westendorp and Kirkwood, 1998). Cummins (1999) raises a number of points in response to this work, related to the extremely uneven distribution of human reproductive success, citing, among other things, the fact that in the 1912 Australian census 50% of the children were the offspring of one in nine of the men and one in seven of the women. Furthermore, 60% of all children died unmarried, and one in nine marriages produced no offspring. He relates these facts to a strong hereditary predisposition to infertility in humans, and strong selection on critical genes on the Y chromosome.

Using Equation 8.8, Seielstad *et al.* (1998) study the effect of polygyny on genetic diversity, and quote a global value of 4.27 for the ratio of the autosomal effective population to the Y-chromosome effective population, calculated by using data from 66 populations. The difference between this figure and the expected value of 4, in the absence of polygyny, is essentially undetectable. They explain this small deviation by pointing out that, although polygyny is practised in two thirds of traditional societies, typically fewer than 20% of males control enough resources to support more than one wife, and often additional wives are sisters of the first wife (*sororal* polygyny) and therefore have identical mitochondrial genomes and share, on average, half of their autosomal genes. A third point is that females in polygynous marriages tend to be less fertile than those in monogamous relationships. However,

the selection on the Y chromosome discussed above may also contribute to this difference (Cummins, 1999).

Looking further back, to early hominids, Hrdy (2000) suggests polyandrous mating for the human–chimpanzee common ancestor. This is perhaps surprising, given that less than 2% of current of human societies are so designated, but is less so if the following special cases are taken into account: shortage of women, extramarital affairs, inability of one man to provide sufficient food, shelter or security, wife-sharing with kin, age-mates and allies, and female sequential mating over a lifetime. She gives as examples the facts that over 60% of Aché men spend some brief time in polyandrous marriages, and that most Aché women have children with two or more men. Nevertheless, her main emphasis is on the variety of mating patterns in primates and modern human foragers, and she advises against assuming that a single breeding system can be applied to all hominid ancestors.

Gage (1998) looks at mortality and fertility models for primates, explicitly comparing *Pan* and modern Swedish and Costa Rican populations, while also, as far as possible, considering prehistoric humans. He found that although delayed maturation is most progressed in humans it is probably a recent development. In addition, the fact that the breeding period is most extended and the rate of aging most rapid in *Pan troglodytes* implies that the contemporary *Pan* demographic system is derived, and thus not a good model for the human–chimpanzee common ancestor. He further points out that the details of the evolution of the reproductive life span will differ in complex ways between equilibrium and non-equilibrium populations, a potentially important point to consider in simulations.

Among the extant apes alone there is an extraordinary diversity of mating patterns. Specifically, breeding groups (adults only) in chimpanzees consist of 6–8 females and 2–3 males; gorillas have a single male to 2–3 females; orangutans have 3–5 females per male, the male being an isolated individual; and, of course, bonobos, famous for their sexual habits, have up to 20 in a group, 4 males to perhaps 16 females! Gibbons are monogamous, as are humans, with variations as discussed above.

8.4 Coalescence and biological ancestry

In contrast to the genetic analysis in Section 8.1, a pedigree documents biological relatedness, i.e. relatedness that is traced back through *both* parents, and results in an exponential increase in the number of an individual's ancestors over time. Figure 8.2 shows the path to the most recent common

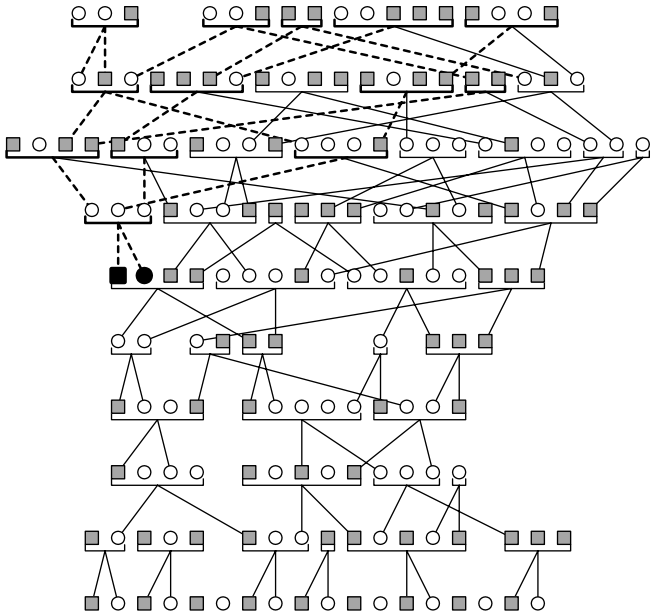


Figure 8.2. The simple genealogy from Figure 8.1, with this time the path from the most recent biological common ancestors to all current individuals highlighted with a thicker dashed line. Note that the highlighted path does not in this case indicate all surviving lineages of the current generation, but just those that lead to the most recent common ancestor pair.

biological ancestors (they will always occur in pairs for a strictly monogamous population) for the sample genealogy of Figure 8.1.

The easiest way to reason about this situation is to imagine a constant population of N couples, where each couple *chooses* two parent couples from the previous generation. Under these circumstances, the probability that the members of a single couple in a population of size N choose the same parents is again $P_1 = 1/N$. However, if they have distinct parents, the probability that they share a grandparent couple is

$$P_2 = R_2(1 - R_4)$$

where R_m, the probability that exactly m distinct *parent choices* are made from the N previous generation couples, is given by

$$R_m = \frac{N^{\underline{m}}}{N^m}, \tag{8.9}$$

where

$$N^{\underline{m}} = m!\binom{N}{m} = \frac{N!}{(N-m)!} \tag{8.10}$$

and is a falling factorial (Graham *et al.*, 1994).[1] In general, the probability that the individuals in the originally chosen couple share an ancestor t generations back is

$$P_t = (1 - R_{2^t}) \prod_{m=1}^{t-1} R_{2^m} \approx \frac{1}{N}\binom{2^t}{2}. \tag{8.11}$$

The coalescence of a pair of lineages in the one-parent case, as described by Equation 8.2, is a very slow process, with an expected coalescence time of N generations. Even when an entire sample is considered, the persistence of distinct lineages is quite strong. Except possibly in the most recent section of the coalescent tree, going back each generation will see the number of lineages either unchanged or reduced by one, and, as one would expect, the likelihood of a coalescence occurring decreases with the number of lineages. The fundamental difference in the two-parent case is, of course, the fact that going back one generation, any lineage that does not coalesce actually bifurcates, and so the number of lineages can increase. In the case of a pair of lineages as analysed above, this rapid proliferation of lineages also means that very quickly 2^t approaches N, and so the large N approximation is not as useful as in Equation 8.2.

The equivalent problem for an entire generation in a population of fixed size N was studied by Chang (1999), who found that for large N the common ancestor is expected in only $\log_2 N$ generations, and unlike in the single-parent analysis, the variance in this case is very small. This very rapid mixing is essentially a result of the exponential increase in the number of ancestors of any given individual, combined with the constraint of fixed population size. The rapidity of this mixing is well illustrated by the example of a population with $N = 5 \times 10^9$ (i.e. approximately equivalent to the size of the current human population). In this case, the biological common ancestor is expected to have occurred a mere 32 generations, or less than 600 years, previously. The limitations inherent in the assumptions underlying models of this type, as discussed in Sections 8.1 and 8.2, are starkly illustrated by

[1] Note that Equation 8.9 is in fact Equation 8.3 expressed in a different notation.

this result: it is apparent that any biological common ancestor of the entire human population must have existed far further back in time. In particular, non-random mating must have acted to keep lineages distinct over periods much longer than 32 generations. Even taking a smaller and potentially more appropriate population of 10 000 breeding individuals, the biological common ancestor is expected only 13 generations, or less than 260 years, back. Yet this two-parent model is based on *exactly* the same assumptions as the one-parent model described in Section 8.1, and if the errors introduced by the underlying assumptions are so significant in the two-parent case they may well be equally problematic in the one-parent case, though less intuitively so owing to the longer time periods involved.

In a constant-population, two-parent model, a sufficient number of generations back all individuals must be either ancestral to the entire current population or on extinct lineages. A second result by Chang (1999) shows that this situation is also reached very quickly: in approximately $1.77 \log_2 N$ generations. Because two-parent genealogies illustrate the mixing of paths of potential autosomal genetic ancestry, the comparison between them and their corresponding one-parent cases raises some interesting questions concerning the relationships between different gene genealogies, and so has implications for the construction and study of species trees. For example, given a tree describing the common mitochondrial ancestry of a population, what degree of biological mixing, and thus potential genetic mixing involving autosomal genes, is likely to have occurred involving the other members of the generation of the mitochondrial common ancestor? Given the N vs. $\log_2 N$ time scales for common ancestry in the one- and two-parent cases respectively, Chang's second result shows that, by the time there is a common mitochondrial ancestor, all individuals who are biologically ancestral to any one current individual will, with certainty, be biological common ancestors. So the answer to the question lies in knowing how many individuals in the mitochondrial common ancestor's generation are on surviving lineages.

The number of surviving lineages in previous generations can be calculated by considering the process of choosing parent couples. Continuing to use the model of a constant population of N couples, where each couple chooses two parent couples from the previous generation, exactly $2N$ choices are made in the first generation, with replacement, from the N potential parents. The probability that exactly m couples from the previous generation are chosen is given by

$$P(m) = \frac{N^{\underline{m}}}{N^{2N}} S(2N, m)$$

where

$$S(2N, m) = \frac{1}{m!} \sum_{k=0}^{m-1} (-1)^k \binom{m}{k} (m-k)^{2N}$$

is a Stirling number of the second kind (Abramowitz and Stegun, 1970; Graham *et al.*, 1994). This comes from there being $N^{\underline{m}}$ ways of selecting the m distinct parent couples (see Equation 8.10) and $S(2N, m)$ ways of assigning the $2N$ individuals making up the offspring couples to these m parent couples, such that each one has at least one child. From the identities

$$\begin{aligned} \sum_{m=1}^{N} N^{\underline{m}} S(k, m) &= N^k \\ \sum_{m=1}^{N} m N^{\underline{m}} S(k, m) &= N^{k+1} - N(N-1)^k \end{aligned}$$

the expected number of previous-generation couples chosen can be calculated:

$$E(N) = \sum_{m=1}^{N} mP(m) = N - N\left(1 - \frac{1}{N}\right)^{2N}. \tag{8.12}$$

Expressing this as a ratio and taking the large N limit gives

$$\lim_{N\to\infty} \left(\frac{E(N)}{N}\right) = 1 - \lim_{N\to\infty} \left(1 - \frac{1}{N}\right)^{2N} = 1 - \frac{1}{\mathrm{e}^2} \tag{8.13}$$

for the fraction of lineages lost when going back a single generation from the present. Going forward in time, this can be seen to be $1 - p(0)$, where $p(0)$ is the probability of zero offspring when modelling the number of offspring as a Poisson distribution with a mean of 2.

Equation 8.12 can be generalised to

$$E(N, k) = \frac{1}{N^k} \sum_{m=1}^{N} m N^{\underline{m}} S(k, m) = N - N\left(1 - \frac{1}{N}\right)^{k} \tag{8.14}$$

for the expected number of couples chosen in k choices from N, leading to a ratio in the large N limit of

$$\lim_{N\to\infty} \left(\frac{E(N, k)}{N}\right) = 1 - \mathrm{e}^{-k/N}. \tag{8.15}$$

Writing A_k for the fraction of the N couples of generation k that are ancestral to the current population, Equation 8.15 can be used to construct the recurrence

$$\begin{aligned} A_0 &= 1 \\ A_{n+1} &= 1 - \mathrm{e}^{-2A_n}. \end{aligned} \tag{8.16}$$

The steady-state fraction of surviving couples in previous generations can then be found by solving Equation 8.16 for any non-trivial fixed points. Specifically, any fixed point is a solution of

$$x = 1 - \mathrm{e}^{-2x},$$

and these are given by

$$x = \frac{1}{2} W(-2\mathrm{e}^{-2}) + 1, \tag{8.17}$$

where $W(x)$ is Lambert's function (Corless *et al.*, 1996), defined as the solution of

$$W(x)\mathrm{e}^{W(x)} = x.$$

This is a multivalued, complex function, but the principal branch, denoted $W_0(x)$, is real for all $x > -1/\mathrm{e}$, and one other branch, $W_{-1}(x)$, is real for $-1/\mathrm{e} < x < 0$. Given Equation 8.17, there are two real solutions for $W(x)$: $W_{-1}(-2\mathrm{e}^{-2}) = -2$, leading to the trivial solution $x = 0$, and $W_0(-2\mathrm{e}^{-2}) \approx -0.4064$, leading to the solution

$$x \approx 0.7968, \tag{8.18}$$

or, approximately, a constant 80% of earlier individuals are ancestral to the current population. By working through the first few terms of this recurrence, as in Table 8.1, it can be seen that the fixed point is approached very rapidly: only four generations back, the ratio already differs from the steady-state ratio by less than 1%.

For completeness, the analogous calculation for the one-parent case proceeds by starting with $k = N$ in Equation 8.15 and generating the similar recurrence

$$\begin{aligned} A_0 &= 1 \\ A_{n+1} &= 1 - \mathrm{e}^{-A_n}, \end{aligned} \tag{8.19}$$

Table 8.1. *First few terms of the recurrence 8.16, and the percentage difference between each term and the non-trivial fixed point* $A^* \approx 0.7968$

n	A_n	$(A_n - A^*)/A^*$
0	1	25.5%
1	0.8647	8.5%
2	0.8226	3.2%
3	0.8070	1.3%
4	0.8010	0.52%
5	0.7985	0.21%
6	0.7975	0.085%
7	0.7971	0.034%
8	0.7969	0.014%

with corresponding fixed-point equation $x = 1 - \mathrm{e}^{-x}$ and solution $x = W(-\mathrm{e}^{-1}) + 1$. At $x = -\mathrm{e}^{-1}$, $W(x)$ has only a single real value of -1 (this is a branch point for W_0 and W_{-1}) and therefore only the trivial solution exists in this case. For this reason, there is no steady-state solution equivalent to the two-parent case. This is as expected, because in the one-parent case the number of prior lineages is strictly decreasing.

Related to the above discussion is the question of how many genes, on average, a biological common ancestor will have contributed to an arbitrary individual in the current generation. Given a steady-state ancestry percentage of approximately 80%, as indicated by Equation 8.18, and a genome of size $g \approx 40\,000$ (International Human Genome Sequencing Consortium, 2001), a biological common ancestor will, on average, be ancestral to $g/(0.8N) \approx 50\,000/N$ genes in the genome of any current individual.

The simulation, *Genie*, is designed to allow direct study of the impact on genealogies of the various issues introduced in this chapter. Full details of the simulation design are presented in the next chapter, but, in brief, genealogies are generated according to the particular settings for population size and changes, breeding patterns and sex ratio, chance of reproduction and expected number of offspring, migration type and rate, selective advantage, and some population-wide external influences. The results are then analysed with respect to both one-parent and two-parent ancestry. The runs described in Chapter 10 focus on a single population (i.e. no migration), and results are compared between various constant-demography cases and those where the demographics vary over time. The runs in Chapter 11 introduce migration,

and those in Chapter 12 introduce the two genetics models used. The first of these models neutral mutations at two loci on each of an autosome, the Y chromosome, and mitochondrial DNA, allowing specification of time-dependent mutation and recombination rates. The second models two alleles with user-controllable degrees of selective advantage and dominance. There is more emphasis on genealogy than on genetics in this volume, but the genetics models included in the simulations allow a substantial degree of sophistication and the output is sufficiently complete to enable quite detailed external analysis if desired. Finally, the results are summarised in Chapter 13 and related to the study of modern human populations.

9 *Simulation design*

Approaches similar to those described in Sections 8.1 and 8.4 allow some quantitative analysis of the issues affecting the generation and analysis of genealogies but are, in general, restricted to highly idealised situations. The program *Genie* is designed to model many of these issues by direct simulation, and thus allow quantitative analysis of their implications in more complex situations, inaccessible to other methods. Nevertheless, practical considerations such as computing power requirements limit the population size and number of generations that can reasonably be simulated on a standard high-end desktop machine to the thousands.

It is worth emphasising that, unlike many of the simulations from the literature referred to in the previous chapter, *Genie* is not explicitly based on coalescent theory. Coalescent-theory-based simulations proceed backwards in time, simulating the *death process* where j lineages are maintained for a length of time t_j, before moving to a state with $j-1$ lineages, and repeating until $j = 1$. Given a specific situation to simulate, i.e. a set of demographic parameters, this process can be completed very rapidly and repeated in order to estimate the probability distribution of the underlying random variable (e.g. the time to the most recent common ancestor) as well as to estimate various point and interval statistics.

In contrast, and similar to the simulations in part I, *Genie* is designed to directly simulate the underlying processes, in a sufficiently generic way to enable a very wide range of demographic regimes, and thus provides a framework for testing the ability of coalescent theory to handle these situations. The simulations run forward in time, and essentially enable the study of several coalescent trees simultaneously. The results are then analysed in the light of coalescent theory, and efficiency is maintained by regularly discarding lineages that can have no impact on the final analysis.

The core of the program allows the simulation of several generations of individuals in up to three independent populations, with parameters controlling the nature of their breeding, offspring, migrations, limiting size, selective advantage and the impact of external, natural disasters. The resulting genealogies are analysed by using both a small sample population and the entire population, and the results compared.

9.1 Parameters

The general parameter settings cover the number of generations to be simulated, the initial sizes of each of three distinct populations (the numbers of males and females are specified independently) and two population size limits along with the generations at which they should be applied. Outside of the specified times, the population is held (approximately) constant at the appropriate limiting value, and between the times the maximum population is determined by exponential interpolation. If only one time limit is provided, the populations are held constant prior to the specified generation, and thereafter are allowed to develop without limitation. The remaining general parameter is the sample size, specified as a percentage of the final population.

The parameters controlling the mating patterns of the simulated individuals apply across all populations. The simulation allows a choice between four different mating patterns: monogamy, polygyny, polyandry and polygynandry. Where appropriate, the size of the mating groups, and the percentage of males therein, can also be specified. In the case of monogamy, no further information is required, whereas for polygyny and polyandry the group size must be specified to determine the required number of females in any mating group. In polygynandry, both the total group size and the percentage of males must be specified to fully determine mating-group composition. The mating parameters may also be made time-dependent, by specifying initial and final generations with their associated settings. Outside the time limits, the settings are constant at the limiting values; between the time limits linear interpolation is used to determine the appropriate values for the size and make-up of the mating groups. Independent male and female *infidelity* probabilities may also be specified, controlling the chance that an individual already a member of one mating group may also be available to participate in another. There is also an option to allow consanguineous matings when the population drops below a certain size.

Related to the mating-pattern settings are those that control the production of offspring. Associated with each female is a chance of reproduction and an average number of offspring. Males are simply randomly chosen until all possible mating groups have been formed. For any particular female, the number of offspring is determined according to a Poisson distribution with the specified mean. It is also possible to specify the sex ratio for the surviving members of the next generation. This feature does not alter the basic 50 : 50 ratio for the sex of individuals at birth, but is implemented as a culling of the new generation if the ratio differs from the specified ratio by more than 10%. If this feature is not used, the birth ratio is maintained. The offspring parameters may be made time-dependent in the same way as the breeding parameters, and, further, may be specified differently for the different populations.

Migration is controlled by specifying a *per* generation probability and percentage involvement for four different types of migration, each with different profiles and consequent side-effects on the populations involved. In the case of a *replacement* migration, the migrants are made up of a number of complete mating groups, and, with equal likelihood, may either replace a group of equal size from the destination population or simply merge into the existing population. *Male-dominated* migrations are where the group of migrants is predominantly male (an arbitrary male to female ratio of 10 : 1 is employed), and a corresponding number of males belonging to the destination population are removed as a result of such a migration (for example, the migrants may be a successful war party). In contrast, *female-dominated* migration models a predominantly female group being taken from a population, with a corresponding loss of source males (for example, if the migrants are taken away as wives or slaves of a successful raiding party). Finally, *itinerants* are modelled as small groups moving essentially independently between populations, with no side-effects at either the source or the destination.

Some of the remaining parameters cover the action of natural disasters, i.e. random population reduction events of various magnitudes. The probability for each of three different sizes of natural disaster, *small* (affecting on average 5% of a population), *medium* (affecting on average 20% of a population) and *large* (affecting on average 50% of a population), can be specified for each population independently. Generation limits between which these probabilities apply can also be specified.

Finally, the program supports simulation of both neutral genetics and selective advantage. Discussion of the models employed is presented in Chapter 12.

As was the case for the species simulation, the random seed used in any particular run is reported, and may be later re-entered to provide an exact reproduction of the earlier run (provided, of course, that the parameter settings are the same). In addition, constraints on the size of the final generation, overall or for each population independently, may be specified. This is especially useful when several runs are being averaged, because, for example, it allows runs where the population dies out to be excluded, or when a particular, but unlikely, outcome needs to be analysed.

9.2 Simulating and analysing a genealogy

In broad terms, the simulation proceeds as follows. Firstly, the initial population is created, including any required marker genes and associated selective advantage, according to the specifications described in the previous section.

Then the simulation loops, each time producing the next generation from the current one, given the parameter settings for mating, offspring and migration. When the final generation is reached, the population is checked against any overall population constraints, and, if satisfactory, the genealogy is analysed. Three kinds of analysis are performed, based on patrilines, matrilines and (optionally) inheritance through both parents, first using a sample of the final population, and then with the entire final population.

In more detail, the generation loop involves the following steps. Firstly, prior to breeding, any required migrations occur and random natural disasters are applied. Then, for each population, as many mating groups as possible are constructed. For each mating group, offspring are generated according to the settings for reproduction chance for each female, and then the number of offspring is determined by using a Poisson distribution with the specified average. Finally, from the full set of offspring so produced, there may be a cull to force the new generation to match the required sex ratio, and the size of the new generation is reduced, if necessary, to conform to the required maximum for this generation.[1] After the new generation has been produced, a number of previous individuals may no longer be ancestral to any members of the current generation. To conserve computer resources, all such individuals are removed from the genealogy, although they remain counted in the historical record of the total population.

Once a genealogy has been produced, a random sample of the current population is taken and used in the subsequent analysis. The lineages of the sample individuals are traced back, firstly along matrilines, thus following mtDNA ancestry, then along patrilines, thus following Y-chromosome ancestry, and optionally through both parents, thus following biological ancestry. Although not strictly genetic, the biological ancestry analysis is important because the lineages so described actually contain all possible gene trees, not just the trees of a highly restricted set of genes as in the other two cases. In addition, the speed of mixing in the biological ancestry case, as discussed in Section 8.4, means that very quickly, with respect to the time of coalescence for any particular gene, the different gene genealogies will differ substantially. The biological ancestry analysis, by quantifying this mixing, provides some information as to the importance of this effect, albeit in a manner that is very hard to decipher. Results from the analysis of the sample population are then compared with an identical analysis based on the entire current population.

The basic information provided by the analysis is the time and location of the most recent common ancestor of each type (i.e. paternal, maternal and biological) for each population and overall, plus a list of migrations occurring within the relevant genealogy. Also provided is the number of extant lineages

[1] This is also the point of application of any required selective advantage (see Chapter 12).

in the respective genealogies at each generation, the breakdown of these lineages by population, and the largest number of current individuals on any one of these lineages. The percentage of individuals in each generation biologically ancestral to at least one member of the current generation is determined, as described in the constant population case by Equation 8.18, as well as a summary of the distribution of the percentage of current individuals that are descended from individual members of the earlier generations. This provides a way to follow the more general analogue of Chang's $1.77 \log_2 N$ result for the number of generations it is necessary to go back before all individuals are either common ancestors or on extinct lineages (Chang, 1999). Finally, the movement of genes through the population is tracked.

The common-ancestor location and lineage-counting analysis proceeds by visiting all ancestors of each member of the current population in turn, incrementing a visit count each time. In any generation, the number of extant lineages is given by the number of individuals in that generation visited at least once, and a common ancestor is any individual whose visit count equals the size of the sample. Obviously, the most recent such individual is the most recent common ancestor. The biological ancestry analysis is the most computationally intensive part of the simulation because of the proliferation of lineages described in Section 8.4, and also is subject to the least variation, and so an option is provided to turn it off for performance reasons.

9.3 Output data and visualisation

The results of the analysis discussed in the previous section are output in four different ways by the simulation. Firstly, the size of each population over time is shown graphically, and information regarding surviving lineages, genes, and migrations may be shown by selecting the appropriate display. Figure 9.1 shows three snapshots of the graphical window for a sample run. In each case, the size of each population over time is apparent, with the current generation at the top of each figure, and the time and location of the paternal, maternal and biological common ancestors, for both the sample population and the full population, are shown by the superimposed symbols. Figure 9.1a indicates the number of individuals in each generation that are ancestral to at least one current individual (in any of the three populations, not just its own population). Figure 9.1b shows the migrations that have occurred, and 9.1c shows the proportion of individuals carrying a particular allele.[2]

[2] Chapter 12 contains a detailed discussion of the presentation of the results of the genetics simulations.

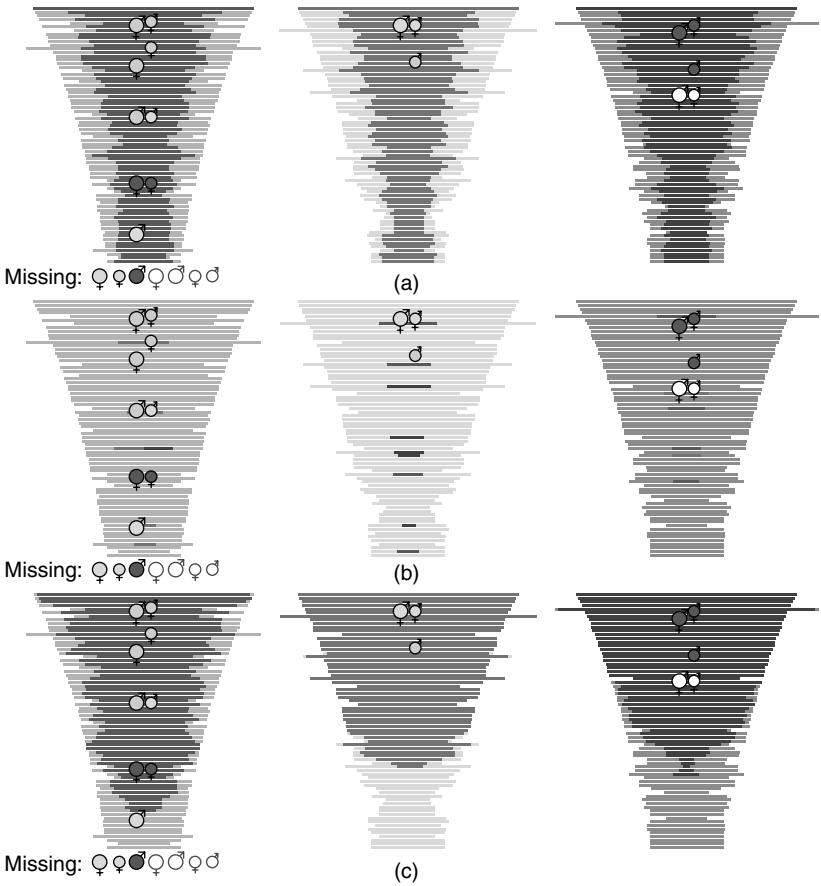


Figure 9.1. A sample run showing three views of the population data, with each of the three populations shaded differently and the current generation at the top of each figure. Common ancestors of each type, for both the sample population and the full population, are shown by the superimposed symbols, with ♂ indicating the paternal common ancestor, ♀ indicating the maternal common ancestor, and the combined symbol ⚥ indicating the biological ancestor. The smaller symbols correspond to the sample population calculations, the larger to the full population, and the shade of each indicates the relevant population, with the overall ancestors indicated by the unshaded symbols. The *missing* ancestors are those that do not exist for the particular run. The internal, darker lines in (a) indicate the proportion of each generation that is ancestral to at least one current individual (in any population). In (b), the darker lines indicate migrants, shaded according to their population of origin. In (c), the darker lines indicate the proportion of individuals carrying a particular allele.

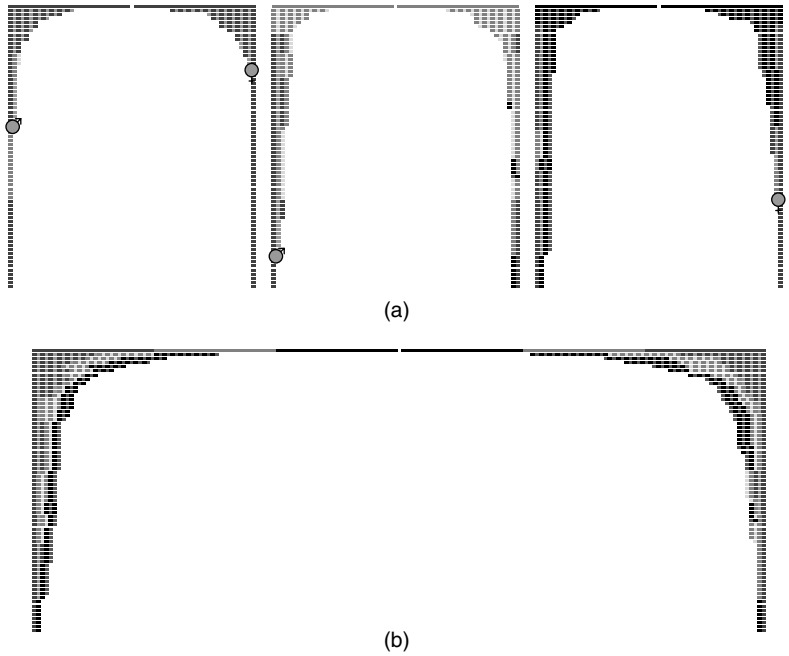


Figure 9.2. Matrilines and patrilines for the populations shown in Figure 9.1. The lineages for each current population are shown separately in (a), and the combined lineages for all populations are shown in (b). The colour indicates the population of a lineage at the particular generation, and common ancestors are indicated as before. The alternating light and dark shading makes it easier to count the number of lineages at any time; the horizontal position of lineages is not significant from generation to generation, i.e. the figure does not indicate the precise continuity of lineages.

The second graphical display shows only the surviving matrilines and patrilines, for each population and overall. These correspond to mtDNA and Y-chromosome lineages, and each lineage is coloured (online) according to the population it is in at the particular generation, thus providing a view of the movement of lineages between populations over time. Figure 9.2 shows the output of this second display for the run shown in Figure 9.1. As in the first graphical display, symbols showing the paternal and maternal common ancestors may be superimposed on the lineage plots, showing the generation they occurred by their position, and coloured according to their population. However, in this case, this only serves to highlight what is already apparent in the figure, since the common ancestor occurs as the point at which the number of lineages drops to one, i.e. the final coalescence, and its population is indicated by the colour of the single lineage at that generation.

These graphical views provide a way to quickly gauge many important aspects of a simulation, but text output is required for a more accurate picture. Because of the amount of data produced by the simulation output and analysis, this is reported in both summary and full form. The text output reports the random seed and all settings used for the run, the ancestry and lineages as calculated by using both a sample and the full population, details of migrations, the number and percentage of individuals in each generation and population that are ancestral to at least one current individual in any population, the relative degrees of current individual ancestry among the earlier generations, and detailed results from the genetics simulation.

In addition, when the graphical output is on display, the underlying data for the particular display are available in text format, for view or export.

Both the summary and full formats show all the settings and common ancestry results, but otherwise the summary format only contains snapshots every ten generations for the lineage and introduced gene results, and fewer samples showing the relative degree of current ancestry. The full format includes these results for all generations and also lists explicitly the migrations in each genealogy. Below is a sample of the summary output showing the final population size, the common ancestry results, the number of migrations in each genealogy, both in total and since the relevant common ancestor, for the sample run in Figures 9.1 and 9.2.

```
##########
## FULL ##
##########
                ==========  ==========  ==========  ==========
                     A           B           C          All
Sample          Y   mtDNA  Bio  Y   mtDNA  Bio  Y   mtDNA  Bio  Y   mtDNA  Bio
                ==========  ==========  ==========  ==========
Population:     30  60     60   30  60     60   30  60     60   90  180    180
Sample size:    30  60     60   30  60     60   30  60     60   90  180    180
Ancestor gen:   40  54     65   8   -      65   -   22     63   -   -      46
Ancestor loc:   A   A      A    A   -      B    -   A      C    -   -      C
Migrations      4   0      19   5   9      20   6   4      20   8   12     21
(since CA)      3   0      0    5   9      0    6   4      1    8   12     5

Total migrations of each type
     Replacement: 21,       Male dominated: 0,    Female dominated: 0,   Itinerants: 0

 LINEAGE INFO FOR THE FULL POPULATION
 ================================================
 Population A: Y               mtDNA            Bio
 ================================================
      70:      30-0-0: 30  (1) 60-0-0: 60  (1) 60-0-0:      60 (1)
      60:      3-0-0:  3   (14) 4-0-0:  4   (23) 28-0-0:     28 (60)
      50:      2-0-0:  2   (24) 1-0-0:  1   (60) 24-24-0:    48 (60)
      40:      1-0-0:  1   (30) 1-0-0:  1   (60) 24-12-18:   54 (60)
      30:      0-1-0:  1   (30) 1-0-0:  1   (60) 28-12-16:   56 (60)
      20:      1-0-0:  1   (30) 1-0-0:  1   (60) 12-14-16:   42 (60)
      10:      1-0-0:  1   (30) 1-0-0:  1   (60) 14-8-14:    36 (60)
 ================================================
 Population B: Y               mtDNA            Bio
 ================================================
      70:      0-30-0: 30  (1) 0-60-0: 60  (1) 0-60-0:      60 (1)
      60:      3-2-0:  5   (18) 1-3-0:  4   (21) 28-30-0:    58 (60)
```

```
     50:      3-1-0:    4  (22) 0-3-0:    3    (34) 24-24-0:     48  (60)
     40:      2-1-0:    3  (28) 0-2-0:    2    (52) 24-12-18:    54  (60)
     30:      1-2-0:    3  (28) 0-1-1:    2    (52) 28-12-16:    56  (60)
     20:      3-0-0:    3  (28) 1-1-0:    2    (52) 12-14-16:    42  (60)
     10:      2-0-0:    2  (28) 1-1-0:    2    (52) 14-8-14:     36  (60)
=================================================================
Population C: Y                mtDNA              Bio
=================================================================
     70:      0-0-30:   30  (1) 0-0-60:   60    (1) 0-0-60:       60  (1)
     60:      2-0-3:    5  (16) 2-0-4:    6    (20) 28-0-34:     62  (60)
     50:      1-1-2:    4  (18) 1-0-3:    4    (24) 24-24-28:    76  (60)
     40:      1-1-2:    4  (18) 3-0-0:    3    (24) 24-12-20:    56  (60)
     30:      0-1-3:    4  (18) 2-0-0:    2    (41) 28-12-16:    56  (60)
     20:      2-0-2:    4  (18) 1-0-0:    1    (60) 12-14-16:    42  (60)
     10:      2-0-2:    4  (18) 1-0-0:    1    (60) 14-8-14:     36  (60)
=================================================================
ALL:           Y                mtDNA              Bio
=================================================================
     70:      30-30-30: 90  (1) 60-60-60: 180   (1) 60-60-60:    180 (1)
     60:      6-2-3:    11 (18) 5-3-4:    12   (27) 28-30-34:    92  (176)
     50:      3-2-2:    7  (46) 1-3-3:    7    (84) 24-24-28:    76  (176)
     40:      2-2-2:    6  (60) 3-2-0:    5    (84) 24-12-20:    56  (180)
     30:      1-2-3:    6  (60) 2-1-1:    4   (101) 28-12-16:    56  (180)
     20:      3-0-2:    5  (60) 1-1-0:    2   (172) 12-14-16:    42  (180)
     10:      2-0-2:    4  (60) 1-1-0:    2   (172) 14-8-14:     36  (180)
```

For the lineage data, the first three numbers give the number of distinct lineages at the given generation in each of the three populations, then the overall number of lineages is shown (simply the sum of the three previous values) and the number in brackets gives the number of current individuals on the largest of the lineages.

The number of individuals in each generation, and the number and percentage of those that are ancestral to at least one current individual in any population, as seen in Figure 9.1a, is shown as follows.

```
DESCENDENT PERCENTAGES
             =============  =============  =============  =============
Generation         A              B              C              All
             =============  =============  =============  =============
70           60/60  (100.0%)  60/60  (100.0%)  60/60  (100.0%)  180/180  (100.0%)
60           28/51  (54.9%)   30/51  (58.8%)   34/51  (66.7%)   92/153   (60.1%)
50           24/35  (68.6%)   24/43  (55.8%)   28/43  (65.1%)   76/121   (62.8%)
40           24/37  (64.9%)   12/27  (44.4%)   20/37  (54.1%)   56/101   (55.4%)
30           28/32  (87.5%)   12/24  (50.0%)   16/32  (50.0%)   56/88    (63.6%)
20           12/16  (75.0%)   14/22  (63.6%)   16/23  (69.6%)   42/61    (68.9%)
10           14/23  (60.9%)   8/15   (53.3%)   14/20  (70.0%)   36/58    (62.1%)
```

This is followed by a table of the degree of ancestry results.

```
RELATIVE DEGREES
             ============================================================
Generations  Pop.  0%     1–11%  12–22%  23–33%  34–44%  45–55%  56–66%  67–77%  78–88%  89–99%  100%
             ============================================================
A
      61–70  524   32.4%  2.4%   5.0%    6.5%    4.2%    1.5%    6.9%    0.4%    3.1%    7.6%    0.0%
      51–60  475   37.7%  0.0%   0.0%    0.0%    0.0%    0.0%    0.0%    0.0%    2.1%    60.2%   0.0%
      41–50  386   39.4%  0.0%   0.0%    0.0%    0.0%    0.0%    0.0%    0.0%    0.0%    60.6%   0.0%
      31–40  336   48.2%  0.0%   0.0%    0.0%    0.0%    0.0%    0.0%    0.0%    0.0%    7.1%    44.6%
      21–30  280   38.6%  0.0%   0.0%    0.0%    0.0%    0.0%    0.0%    0.0%    0.0%    0.0%    61.4%
      11–20  237   42.6%  0.0%   0.0%    0.0%    0.0%    0.0%    0.0%    0.0%    0.0%    0.0%    57.4%
       1–10  211   31.8%  0.0%   0.0%    0.0%    0.0%    0.0%    0.0%    0.0%    0.0%    0.0%    68.2%
```

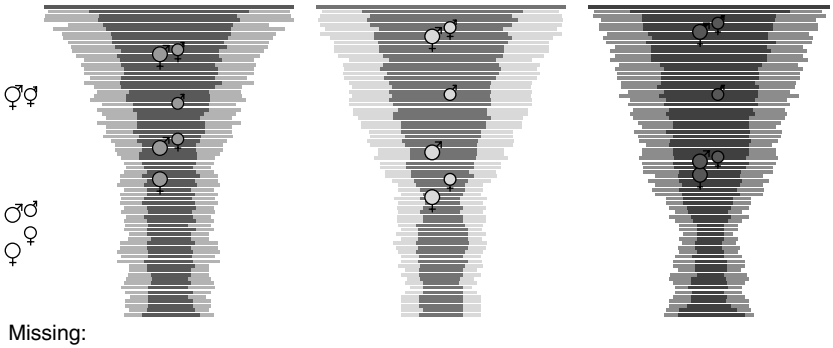


Figure 9.3. Graphical display of the average of ten runs with the same settings as in Figure 9.1. Only the generation of the common ancestors is indicated by the symbols in this case, with the overall common ancestors drawn in the left margin. The darker lines indicate the average survivor percentage for each population.

In this case, each row of results is cumulative over ten consecutive generations and the columns give the percentage of individuals in those generations that are ancestral to the indicated percentage range of current individuals. The percentage that are on extinct lineages, i.e. the 0% column, and the percentage that are common ancestors, i.e. the 100% column, are both given exactly. As expected from the discussion in Section 8.4, before too long there are no individuals that are ancestral to some, but not all, of the current generation. Specifically, by generation 30, i.e. 40 generations back, all individuals are either common ancestors (61.4% of individuals between generations 21 and 30) or on extinct lineages (38.6%). This is somewhat further back than predicted by Chang (1999) for the constant population case of $1.77 \log_2 N \approx 13.3$ generations for $N = 180$, primarily as a result of the migrations.

The simulation output is slightly different when multiple runs are being averaged. For the graphical population display, as shown in Figure 9.3, it is no longer possible to indicate location information for the average common ancestors, because they may be distributed over all three populations. Therefore, only the average generation of each common ancestor is shown, with the symbols indicating the single-population common ancestors superimposed over the relevant population, and the symbols indicating overall common ancestry drawn in the left margin.

The text output is also slightly different for average runs. Although it is possible to output the full data for each run in the average, the default is simply to output the basic ancestor analysis for each of these, followed by a detailed summary. Contained in this summary are the average generations

for the different kinds of common ancestor, their standard deviation, and the percentage of runs where each ancestor comes from each of the different populations, or is not present at all. In each case, only those runs where a common ancestor was found are included in the generation average. Also shown is the frequency with which the common ancestors of the different kinds actually occur in the same population, how often the sample analysis found the true common ancestor for each type, and the average number of migrations in each tree. Here is the ancestor and migration summary for the runs shown in Figure 9.3.

```
## SUMMARY ##
Rejected 10 of 20
                 ============   ============   ============   ==============
                       A              B              C              All
Sample           Y     mtDNA  Bio     Y      mtDNA  Bio     Y      mtDNA  Bio     Y       mtDNA   Bio
                 ============   ============   ============   ==============
Population:      28.1  56.5   56.5    28.2   56.7   56.7    26.4   54.5   54.5    82.7    167.7   167.7
############
## SAMPLED ##
############
Ancestor gen:    48.56 40.25  60.70   50.67  31.78  65.40   50.50  36.50  66.20   24.83   19.00   50.40
Std. dev.:       14.13 15.45  14.66   18.23  19.50  0.70    12.99  21.19  1.48    8.98    9.66    12.72
Ancestor at A:   50.0% 40.0%  90.0%   20.0%  30.0%  10.0%   20.0%  20.0%  0.0%    10.0%   30.0%   50.0%
Ancestor at B:   30.0% 10.0%  10.0%   50.0%  40.0%  90.0%   0.0%   20.0%  10.0%   30.0%   20.0%   20.0%
Ancestor at C:   10.0% 30.0%  0.0%    20.0%  20.0%  0.0%    60.0%  40.0%  90.0%   20.0%   20.0%   30.0%
No Ancestor:     10.0% 20.0%  0.0%    10.0%  10.0%  0.0%    20.0%  20.0%  0.0%    40.0%   30.0%   0.0%
Ancestor Location Matches
Y/mtDNA:         4 of 7 (57.1%)       7 of 9 (77.8%)       4 of 7 (57.1%)       3 of 5  (60.0%)
mtDNA/Bio:       5 of 8 (62.5%)       4 of 9 (44.4%)       4 of 8 (50.0%)       5 of 7  (71.4%)
Y/Bio:           4 of 9 (44.4%)       5 of 9 (55.6%)       6 of 8 (75.0%)       2 of 6  (33.3%)

Migrations       2.8    3.0   15.5    4.2    3.6   15.5    2.0    3.3   14.0    6.0     6.5     18.5
(since CA)       1.5    2.5   1.1     2.2    3.0   0.6     0.9    2.7   0.4     5.2     6.3     4.6

############
##  FULL  ##
############
Ancestor gen:    38.17 31.13  59.30   37.78  27.75  63.00   35.50  32.63  64.80   23.50   15.33   50.40
Std. dev.:       16.08 15.32  14.37   16.02  18.11  4.94    17.53  16.11  0.79    9.18    6.09    12.72
Ancestor at A:   20.0% 50.0%  80.0%   20.0%  40.0%  0.0%    10.0%  20.0%  0.0%    10.0%   30.0%   50.0%
Ancestor at B:   30.0% 10.0%  20.0%   40.0%  30.0%  90.0%   20.0%  20.0%  0.0%    30.0%   20.0%   20.0%
Ancestor at C:   10.0% 20.0%  0.0%    30.0%  10.0%  10.0%   50.0%  40.0%  100.0%  20.0%   10.0%   30.0%
No Ancestor:     40.0% 20.0%  0.0%    10.0%  20.0%  0.0%    20.0%  20.0%  0.0%    40.0%   40.0%   0.0%
Ancestor Location Matches
Y/mtDNA:         4 of 5 (80.0%)       6 of 8 (75.0%)       6 of 8 (75.0%)       4 of 4  (100.0%)
mtDNA/Bio:       5 of 8 (62.5%)       4 of 8 (50.0%)       4 of 8 (50.0%)       5 of 6  (83.3%)
Y/Bio:           3 of 6 (50.0%)       5 of 9 (55.6%)       5 of 8 (62.5%)       4 of 6  (66.7%)

Migrations       4.7    4.0   15.5    4.9    4.7   15.5    3.2    3.6   14.9    7.4     7.4     18.5
(since CA)       3.6    3.6   1.4     3.5    4.3   1.1     2.1    2.8   0.5     6.7     7.2     4.6

Sample/true ancestor matches
Y:               5 of 9 (55.6%)       4 of 9 (44.4%)       2 of 8 (25.0%)       4 of 6  (66.7%)
mtDNA:           5 of 8 (62.5%)       7 of 9 (77.8%)       5 of 8 (62.5%)       5 of 7  (71.4%)
Bio:             5 of 10 (50.0%)      2 of 10 (20.0%)      2 of 10 (20.0%)      10 of 10 (100.0%)

Total migrations of each type
        Replacement: 19.9,       Male dominated: 0,     Female dominated: 0.0,        Itinerants: 0.0
```

The lineage, ancestry and descendants information is as for the single-run case, except that average values are shown and the size of the largest lineage is no longer reported.

10 *Simulating a single population*

Given the wide variety of parameters available to the simulation, trying to isolate and understand the important effects of each is of obvious importance in determining exactly how it can be applied to particular situations of interest. For this reason only single-population simulations are employed throughout this chapter, initially keeping as close as possible to the simple constant population theory outlined in Section 8.1, and then allowing the parameters controlling population size, sex ratio, fertility rate and breeding pattern to change in controlled ways.

Basic coalescent theory (Kingman, 1982b; Hudson, 1990) describes lineage merging in terms of the effective population size, N, and can thus be applied to populations with many different mating patterns and sex ratios by using an appropriate value for N. Extensions to varying population size have also been investigated (Harpending *et al*., 1998; Tavaré *et al*., 1997), but the variation in these models is quite limited in nature. The simulation *Genie* allows study of a much wider range of problems, but it must first be demonstrated that it agrees with the theory in these simple limits.

10.1 Constant demographics

The theory outlined in Section 8.1 is for a population of constant size, and therefore the first set of simulations focus on this situation specifically. However, as the following results show, this still allows investigation of many interesting and important features of the simulations. Specifically, constant population simulations are presented with various fertility rates, mating patterns and chances, and degrees of infidelity, and with different restrictions on consanguineous mating. The runs cover the role of these basic parameters and thus provide a foundation for the more complicated scenarios to follow.

10.1.1 Results for different fertility rates

The simulation does not allow an exact reproduction of the theoretical case described in Section 8.1, because even with each female having a 100%

reproduction chance, and number of offspring Poisson distributed with average 2, random effects will cause the total population to fluctuate and the sex ratio to become imbalanced, leading very rapidly to extinction. The closest approximation that can be made is to use the population limiting factors, with no restriction on mating between related individuals, and a total fertility rate (TFR) that ensures sufficient offspring each generation.

The lowest sustainable TFR for constant population runs in the simulation depends on the population size: the larger the population, the closer the simulation can come to the theoretical case. For example, with a population of 1000 males and 1000 females, an average fertility rate of 2.2 offspring per female results in a sustainable constant population simulation, but when the total population is kept to only 200 individuals, as is the case for most of the simulations in this chapter, a fertility rate of approximately 2.5 is required. Of course, as the population size increases, the random mating assumption becomes less realistic, and so simply increasing the population size does not necessarily overcome this limitation.

The following runs were designed to find the most appropriate fertility rate setting for a population with a constant size of 200 individuals, consisting of equal numbers of males and females, by increasing the fertility rate from 2.5 to 12 while keeping all other parameters constant. The results for the time to the most recent common ancestor and the steady-state percentage of prior individuals ancestral to the current population are shown in Table 10.1.

Given the population-limiting factors in the simulation, the theoretical expectation is that, once the TFR is sufficiently large to ensure the limiting value is met each generation, different TFRs should not lead to any difference in the coalescent analysis (the null hypothesis). Although the individual common ancestor generations are not normally distributed, the Central Limit Theorem means that the average common ancestor generation (certainly for a simulation average of 100 or more) is essentially a normally distributed continuous variable, and therefore, given equal variance, an ANOVA can be used to see whether there is any significant difference in the means in Table 10.1. The results of the ANOVA are a P value of 0.61 for the paternal common ancestor analyses, 0.86 for the maternal common ancestor analyses, and 0.48 for the biological common ancestor analyses. Thus all three values strongly support the null hypothesis.

This conclusion is further strengthened by carrying out a t-test to compare the extreme cases of TFR = 2.5 and TFR = 12, where the resulting P values of 0.38 for the paternal ancestor comparison, 0.60 for the maternal ancestor comparison and 0.49 for the biological ancestor comparison all comfortably support the null hypothesis of same population.

Table 10.1. *The time to the most recent common ancestor, and the steady-state percentage of prior individuals ancestral to the current population, for constant population simulations with* N = 200*, a 50:50 sex ratio, and six different values for the total fertility rate (TFR)*

An ANOVA of the means indicates no significant effect on the results from the fertility rate setting, so a TFR of 4 was chosen as the standard value for maintaining both a constant population and proximity to the theoretical model described in Section 8.1.

	Generations back to common ancestor (s_{N-1})						
TFR	Paternal		Maternal		Biological		Survivors
2.5	190.5	(97.1)	213.7	(96.0)	7.96	(0.28)	79.6%
2.75	201.9	(105.7)	196.9	(101.6)	8.02	(0.35)	79.8%
3.0	181.8	(88.1)	195.5	(100.4)	7.97	(0.30)	80.2%
4.0	191.2	(105.3)	200.6	(99.1)	7.96	(0.28)	79.8%
8.0	188.6	(98.5)	207.7	(107.4)	7.95	(0.36)	79.4%
12.0	204.6	(112.4)	206.1	(107.7)	7.93	(0.33)	79.3%

Given these results, a TFR of 4 was chosen as a standard value for maintaining both a constant population and proximity to the theoretical model described in Section 8.1.

The next comparison of interest is that of the TFR = 4 case against the theoretical expectation of a mean time to the common ancestor of $2N$, with variance $1.16N$. Applying a z-test results in P values of 0.42 for the paternal common ancestor generation and 0.96 for the maternal common ancestor generation, indicating strong support of the hypothesis of agreement with the theoretical results.

For all TFR values, approximately 80% of prior individuals were ancestral to at least one member (and, indeed, in most cases all) of the current population, as predicted by Equation 8.18 and the work of Chang (1999). An extremely rapid mixing of biological lineages was seen, with very little variance. However, the average time of just under 8.0 generations back was significantly longer than the predicted average of $\log_2 N \approx 7.64$ generations. With a somewhat larger population size of $N = 2000$, the biological common-ancestor generation average was 11.6 generations back, compared with the theoretical value of 11.0, and ten runs with a population of 20 000 had the biological common ancestor 15 generations back every time, compared with $\log_2 N \approx 14.3$ generations. This difference, though significant, is small, is consistent with the simulations in Chang (1999), and can be explained by the asymptotics.

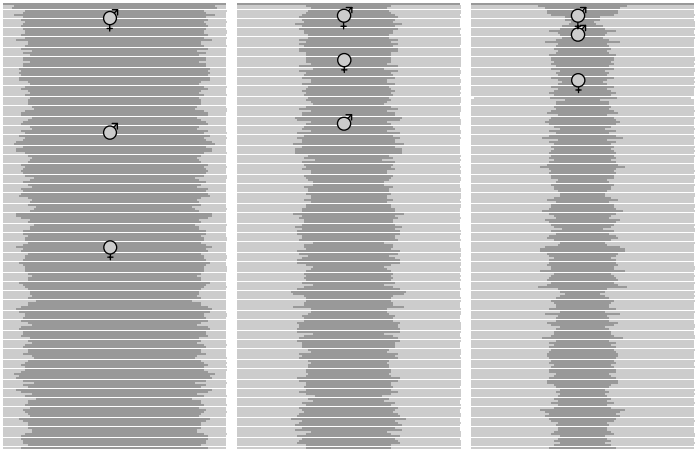
The average times to the common ancestors of each type as calculated from a 5% sample population were always a little less than the corresponding true values. With the small population of 200 individuals, and the 5% sample size, in more than 60% of cases the true paternal common ancestor was found, increasing to around 90% of cases for the maternal common ancestor. The improvement for the maternal ancestor results from the fact that current males can still be included in the maternal sample (since males carry mtDNA but do not pass it on) and thus the maternal ancestry sample size for a population of 100 individuals was 10, rather than only 5. These values are consistent with the expected $(m-1)/(m+1)$ value for $m=5$ and $m=10$, respectively (Saunders *et al.*, 1984). However, the true biological common ancestor was found by the sample in less than 10% of cases, further indicating the difference in the lineage-mixing for this case.

10.1.2 Results for different reproduction chance

In each of the simulations discussed in Section 10.1.1, 100% of females were able to mate successfully (although the offspring distribution ensured that not all had children). The simulations in this section were designed to compare the above situation of 100% chance of reproduction and an average of four offspring per female, with one where the chance of reproduction was only 50% per female but the average fertility was doubled to eight offspring, as well as the more extreme situation of a chance of reproduction of 33% per female and an average fertility of 12 offspring. Each of these situations led to the same expected total number of offspring, but obviously the effective population was quite different in each case, and subsequently the lineage mixing profiles were also very different.

A single-run example is shown in Figure 10.1. In a strictly monogamous population, reducing the percentage of females that can reproduce results in an equivalent reduction in the number of males that can reproduce. So the effective population for both sexes changes in proportion to the reproduction chance.

The maternal and paternal common ancestors show significant variability across these three runs, indicating the difficulty in applying general rules to particular cases when the variance is large. Nevertheless, the tendency for the time to the most recent common ancestor to be less in the higher-fertility/lower-chance runs was apparent, resulting from the much increased likelihood that any two individuals shared a parent in these cases. In fact, just ten generations back there were only two surviving paternal and maternal lineages in the 12/33% case, compared with eight for both in the 4/100% case.



(a) 4/100% (b) 8/50% (c) 12/33%

TFR	Chance	Effective population	Generations back to common ancestor			Survivors
			Paternal	Maternal	Biological	
4	100%	100	58	110	7	79%
8	50%	50	54	26	6	40%
12	33%	33	14	35	6	28%

Figure 10.1. Three single runs with different reproduction chance/TFR pairs. Note the wide variation in common-ancestor generation (for the non-biological ancestor cases), and the scaling of the survivor percentage with the effective population.

In contrast to the single-parent result, the $\log_2 N$ prediction for the number of generations to the most recent biological common ancestor was closely matched in all three cases. This is a result of the much lower variance for the process of biological ancestry.

Average results over 100 runs with these three configurations are shown in Table 10.2. In all cases the average common-ancestor generation was consistent with the theoretical prediction of twice the effective population. Figure 10.2 shows the number of paternal lineages for the 4/100%, 8/50% and 12/33% cases over the final 300 generations. The maternal lineage profile was almost identical since the effective population was the same for males and females, and an average of 100 runs is sufficient to ensure close proximity to the limiting behaviour.

Consistent with the rapid lineage mixing that leads to a biological common ancestor after only a few generations, the number of biological lineages settled

Table 10.2. *Average results with three different reproduction chance/TFR pairs*

		Generations back to common ancestor (s_{N-1})						
TFR	Chance	Paternal		Maternal		Biological		Survivors
4	100%	191.3	(105.3)	200.6	(99.1)	7.96	(0.28)	79.8%
8	50%	99.0	(51.7)	106.9	(64.3)	6.88	(0.46)	40.3%
12	33%	64.1	(37.8)	67.9	(39.5)	6.29	(0.50)	26.7%

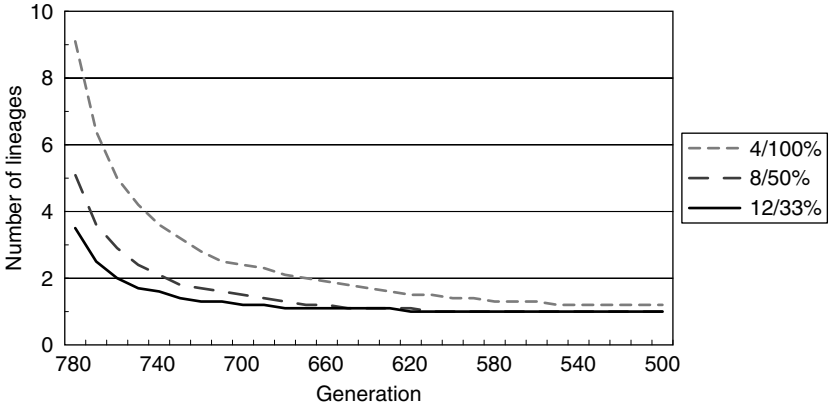
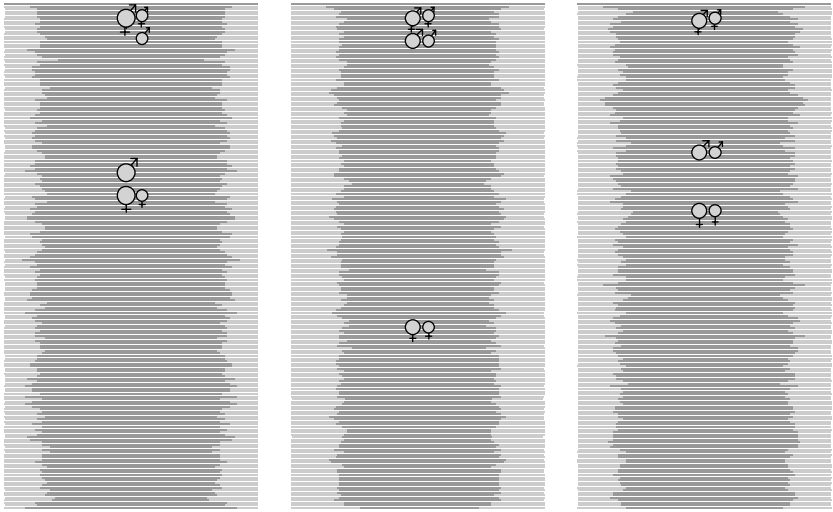


Figure 10.2. Paternal lineages for generations 780 to 500 with three different reproduction chance/TFR pairs. The maternal lineage plot is similar.

down to the steady-state value given by Equation 8.18, scaled by the effective population, after only five or six generations in all cases.

10.1.3 Results for different mating patterns

For all results shown so far the mating pattern has been monogamy, and thus the paternal and maternal lineage analyses have essentially been parallel examples of the same process. Figure 10.3 shows three sample runs, each with a different mating pattern. For the non-monogamous runs, the effective population is now different for males and females, leading to a significant difference in the times to the common ancestors for the different sexes. Each of these runs had a constant population size of 100 individuals, so the male



(a) *MF* (b) *MFFF* (c) *MMF*

Mating group	Per cent male	Generations back to CA			Survivors
		Paternal	Maternal	Biological	
MF	50%	67	76	6	73.2%
MFFF	25%	15	128	6	62.4%
MMF	67%	59	82	7	68.1%

Figure 10.3. Single runs with three different mating patterns and population adult sex ratios, for 100 individuals over 400 generations. The different mating patterns are monogamy (*MF*), polygyny with three females per male (*MFFF*) and polyandry with two males per female (*MMF*).

effective population was equal to the male percentage, as shown in the table at the bottom of Figure 10.3. Furthermore, because the population sex ratio matches that of the mating groups, all individuals were able to mate and the female effective population was thus 100 minus the male figure. No clear difference was seen in the biological common ancestor results for these runs.

When extended to averages over 100 simulations, for a population of 200 individuals, the results were as shown in Table 10.3. These results show very clearly the different ancestral features of the different mating patterns. The adult population sex ratios for the monogamy, polygyny and polyandry runs were such that they matched the relevant mating-group sex ratio, but for the polygynandry runs the sex ratio was 50% and thus did not match

Table 10.3. *Average results of 100 runs with four different mating patterns*

The numbers of males and females in mating groups are as indicated by the string of *M*s and *F*s, and the adult population sex ratio is also shown.

Mating group (% male)	Generations back to common ancestor (s_{N-1})						Survivors
	Paternal		Maternal		Biological		
MF (50%)	191.3	(105.3)	200.6	(99.1)	7.96	(0.28)	79.8%
MFFF (25%)	104.1	(55.0)	292.2	(155.2)	7.39	(0.53)	68.3%
MMMF (75%)	307.6	(145.8)	108.7	(69.3)	7.42	(0.52)	67.9%
MMFFFF (50%)	96.7	(55.1)	196.6	(108.8)	7.41	(0.49)	56.1%

the sex ratio in the mating groups. Therefore, for these runs, only half the males in the population were able to participate in mating groups. Neither the paternal nor the maternal results showed a significant departure from the relevant theoretical predictions from coalescent theory.

The rate of reduction of both paternal and maternal lineages is shown in Figure 10.4. In each case, the rate of lineage-merging was directly related to the effective population. This was most clearly demonstrated by the polygynandry case, where the male effective population matched that of the polygyny case, and the female effective population was as for the monogamy case. The corresponding curves in the figure very closely match each other, while remaining well separated from the curves for the other mating patterns.

The survivor percentages, as shown in Table 10.3, varied significantly across the different cases, but the biological common ancestors remained very close in time and indeed were indistinguishable between the three polygamous cases. Using Equation 8.8 to calculate the effective population relevant to the inheritance of an autosomal gene gives $N_{\text{au}} = 200$ for monogamy, $N_{\text{au}} = 150$ for polygyny and polyandry, and $N_{\text{au}} = 133$ for polygynandry. The steady-state percentages are not precisely in the same ratios as these figures, because the calculation in Section 8.4 that leads to Equation 8.18 is based on monogamous couples and cannot be used. Nevertheless, it still gives a broad indication of the relationship between them.

A sample population of size 5% was taken and analysed in all runs. For the male sample in the polygyny run, this meant that only two individuals were sampled. Nevertheless, in 44% of cases the most recent common paternal ancestor of these two individuals was the most recent common paternal ancestor of the whole population – better, in this case, than the prediction of 33% from the $(m-1)/(m+1)$ rule. Because of the greater percentage of males in the polyandrous population, it is to be expected that for these runs

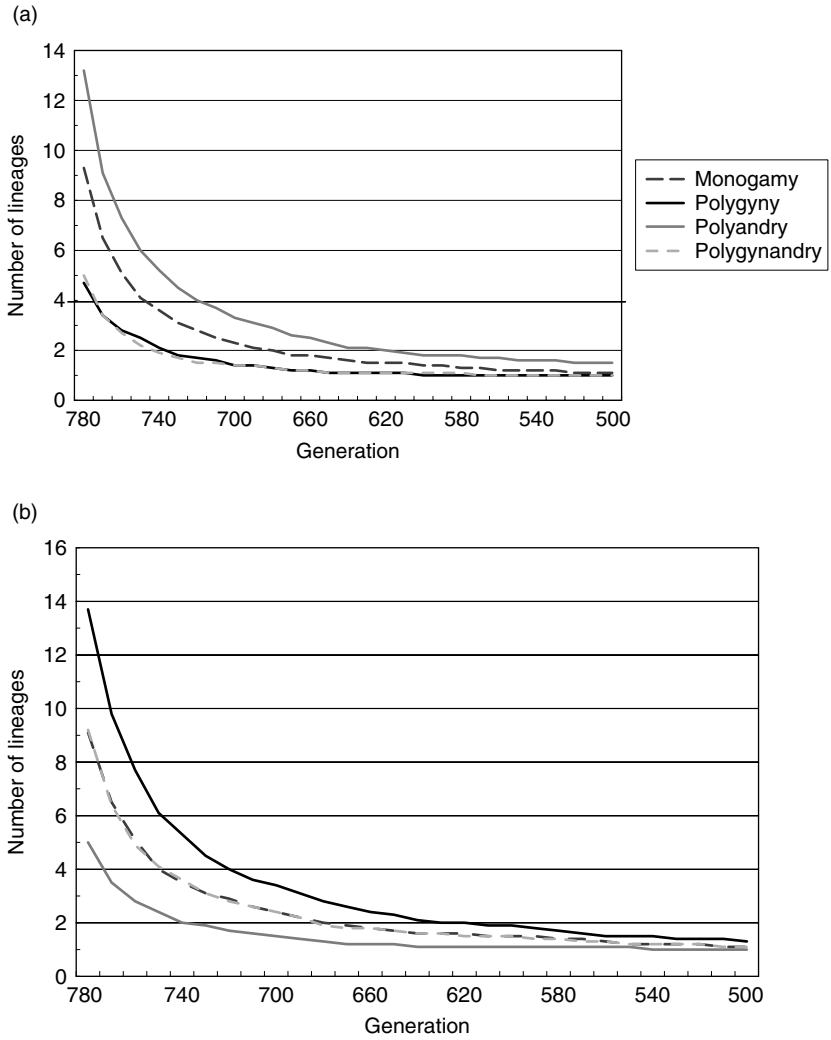


Figure 10.4. Average numbers of surviving paternal (a) and maternal (b) lineages over the final 300 generations for four different mating patterns: monogamy (*MF*), polygyny (*MFFF*), polyandry (*MMMF*) and polygynandry (*MMFFFF*).

the male sample will have greater accuracy than in the other cases, and this was indeed the case. In fact, for 79% of the polyandry runs, the most recent common paternal ancestor of the male sample individuals was the most recent common paternal ancestor of the whole population, compared

with only 74% of runs where the most recent maternal common ancestor was the same for both the sample and full populations. As in the earlier results, the most recent biological common ancestor of the sample individuals was only rarely the most recent biological common ancestor of the entire population, but because of the extremely rapid mixing the sample common ancestor generation was never further than a few generations from the true value.

10.1.4 Restricting consanguineous matings

In all of the above runs, an individual's choice of mate was purely random. One way in which this assumption is violated in real populations is by geographical or social separation preventing, or at least minimising, interbreeding. This situation is studied in the next chapter. Another important way in which this assumption is violated is through the restriction of mating between close relatives, in particular siblings and cousins. (Because generations in the simulation are discrete, there is never any possibility of mating between parents and their offspring.) In the following runs, only if the population dropped to 15 individuals or fewer were consanguineous matings allowed, because in such cases the only alternative was dying out.

Simulations were run for the three different mating chance cases in Section 10.1.2, but with consanguineous matings disallowed and a smaller population of 100 individuals. The results in Table 10.4 show that it was almost impossible for the TFR 12/33% case to survive for 400 generations. In fact, in none of the 100 runs did this population survive for even 60 generations, thus demonstrating the effects of random fluctuations and the dependence on a large enough population size to ensure survival. When run instead with a constant population of 200 individuals, a population with these settings was seen to be far more likely to survive. Table 10.5 shows the average results over 100 runs for the same cases as shown in Table 10.2, except that here the results for consanguineous matings both allowed and disallowed are included. In all cases, the survivor percentages decreased when consanguineous mating was restricted, and there was a related significant difference in the biological ancestor figures. Furthermore, as the chance of reproduction decreased, the effect of the consanguineous mating restriction became more apparent in the single-parent common ancestor generations. When a *t*-test was applied to the single-parent ancestry results, no effect was indicated for 100 run averages in the 100% mating chance cases, for either monogamy or polygyny, but P values of 0.038 and 0.029 for the paternal and maternal common ancestor generations in the 12/33% case indicate a

Table 10.4. *Average of 100 smaller population runs, with the three different TFR/mating chance pairs from Section 10.1.2, but here with consanguineous mating disallowed for populations greater than 15 individuals*

All runs were over 400 generations.

		Generations back to common ancestor (s_{N-1})						
TFR	Chance	Paternal		Maternal		Biological		Survivors
4	100%	94.9	(60.5)	88.2	(45.8)	6.56	(0.50)	73.9%
8	50%	37.3	(21.7)	37.0	(18.8)	5.26	(0.52)	33.7%
12	33%	Population died out in all instances						

Table 10.5. *Various TFR/reproduction chance runs, with consanguineous mating both allowed and disallowed*

The light grey background highlights the cases where consanguineous mating is disallowed; results for the other cases are simply taken from Table 10.2 and included here for ease of comparison.

		Generations back to common ancestor (s_{N-1})						
TFR/chance	Cons. mating	Paternal		Maternal		Biological		Survivors
Monogamy								
4/100%	Yes	191.3	(105.3)	200.6	(99.1)	7.96	(0.28)	79.8%
4/100%	No	193.4	(109.4)	195.2	(101.7)	7.80	(0.40)	76.6%
8/50%	Yes	99.0	(51.7)	106.9	(64.3)	6.88	(0.46)	40.3%
8/50%	No	81.7	(40.2)	95.0	(55.4)	6.64	(0.48)	36.8%
12/33%	Yes	64.1	(37.8)	67.9	(39.5)	6.29	(0.50)	26.7%
12/33%	No	54.0	(29.9)	55.9	(37.4)	5.80	(0.45)	22.9%
Polygyny (*MFFF*)								
4/100%	Yes	104.1	(55.0)	292.2	(155.2)	7.39	(0.53)	68.3%
4/100%	No	99.2	(59.6)	273.8	(149.1)	7.16	(0.39)	65.2%

significant effect at this level. The 8/50% results were less clear, with a P value for the paternal common ancestor generation of 0.16, indicating no effect, but for the maternal common ancestor generation the P value was 0.009, strongly suggesting a significant effect. Ideally, these two measures should be in agreement.

There can be no doubt that disallowing consanguineous mating does have a genuine influence on the results, because by reducing the size of the pool

of potential mates it has an impact on the effective population. As population size increases, the effect becomes less important, but, as discussed earlier, larger populations are less realistic because the random mating assumption is less applicable. What these results showed is that, for 100 run averages, in the 100% reproduction chance cases any real effect was swamped by the random variation. However, if the above reasoning is true, the effect should be apparent if more runs are averaged and thus a more accurate estimate of the average common ancestor generation obtained.

In order to resolve this issue better, the polygyny example was run again, this time averaging over 1000 simulations. When consanguineous matings were allowed, a paternal common-ancestor time of 100.2 generations back, with standard deviation 52.9, was found, matching almost exactly the theoretical prediction. In contrast, the maternal common-ancestor time of 284.6 generations back, with standard deviation 135.7, is significantly more recent than the theoretical prediction of 300 generations. This is an indication of the departure of the simulation from the ideal model in a simulation with small N, and was most apparent for maternal case because of the increased sensitivity to the sex ratio given the larger population of females in the polygyny simulation. However, in the case where consanguineous matings were disallowed, both ancestor times were significantly more recent than predicted by the ideal model: 94.4 generations (with $s_{N-1} = 50.5$) for the paternal ancestor, and 262.2 generations (with $s_{N-1} = 129.3$) for the maternal ancestor. Testing these against the theoretical expectations with a z-test gives P values of 0.001 and 10^{-13}, respectively. Using a t-test to compare these results with each other gives P values of 0.012 for the paternal-ancestor generation case and 0.00016 for the maternal-ancestor generation case, very strongly indicating a statistically significant difference in the results.

10.1.5 Infidelity

The simulations described in Section 10.1.3 compared fundamentally different mating patterns. However, the infidelity settings provide a way to simulate a genealogy in which there is a degree of corruption of an otherwise pure mating pattern. The effect of introducing *infidelity* (or equivalently, serial mating) into a simulation was examined by running some $N = 200$ monogamy simulations with the male infidelity rate set to 25% and the female infidelity rate set to 10% (no consanguineous mating allowed). The common-ancestor analysis gave an average time to the most recent paternal common ancestor of 136.2 generations back (standard deviation 70.5), compared with 166.7 generations back

(standard deviation 79.7) for the most recent maternal common ancestor and 7.30 generations back (standard deviation 0.46) for the most recent biological common ancestor. All three of these values were significantly different from the pure monogamy average results shown in Table 10.5, where the most recent paternal common-ancestor generation was 193.4 generations back, the most recent maternal common-ancestor generation 195.2 generations back and the most recent biological common-ancestor generation 7.80 generations back. In addition, the historical survivor percentage dropped by more than 10%, from 76.6% down to 68.5%. The impact was greatest for the paternal genealogies because males had the largest likelihood of multiple partners.

10.2 Varying demographics

The role of the basic parameters in various constant-population simulations was covered in Section 10.1. The simulations studied in this section are those in which the parameters controlling the population size, breeding pattern and sex ratio changed over time. Drawing on the previous results, particular emphasis was placed on situations where there was an overlap between the average times of the most recent common ancestors for the constant demography runs and the time of the demographic changes in the varying runs. All simulations were configured to finish with a population of 200 individuals.

10.2.1 Variation in population size

The first set of runs involved breaking the constant population size restriction in various ways: specifically, long-term exponential growth, brief periods of exponential growth, bottlenecks, and random fluctuations. Initially, the simulations were restricted to the simplest case of monogamy with TFR $= 4$ and 100% reproduction chance. Average results for these various possibilities are shown in Table 10.6.

If we look first at the basic profile averages, it is clear that the constant population case and the early exponential growth cases were very similar, and of the other three, the pure exponential growth and middle exponential growth cases were closest. This seems quite reasonable given the average profiles for the exponentially growing cases, shown in Figure 10.5. The degree of similarity between the constant-population case and the early exponential growth case for the two single-parent ancestor calculations is perhaps a little surprising, because there would certainly have been some runs (out of the 100 averaged) where the common ancestor fell in the region in which these

Table 10.6. *Table of average results for various population size possibilities and a monogamous mating pattern*

Consanguineous matings were disallowed for populations >15 in all cases. Bottleneck times were between generations 250 and 300 for early, generations 500 and 550 for middle, and generations 650 and 700 for late (out of a total of 800 in all cases). The early exponential growth runs had 300 generations of growth followed by 300 generations of constant population size; in the middle exponential growth runs there were only 200 generations of constant population size at the end, and the late exponential growth runs had exponential growth from a small population only over the last 200 generations (see Figure 10.5).

		Generations back to common ancestor (s_{N-1})					
Basic type	Modification	Paternal		Maternal		Biological	
Basic profiles							
Constant	None	193.4	(109.4)	195.2	(101.7)	7.80	(0.40)
Exponential	None	148.9	(61.6)	157.0	(70.5)	7.74	(0.44)
Exponential	Early	194.5	(85.9)	201.8	(89.5)	7.83	(0.38)
Exponential	Middle	163.3	(66.9)	172.0	(64.6)	7.86	(0.35)
Exponential	Late	111.4	(40.5)	119.1	(43.2)	7.75	(0.44)
Bottlenecks							
Constant	Early	199.3	(99.7)	207.4	(103.5)	7.78	(0.42)
Constant	Middle	169.9	(66.8)	171.5	(65.3)	7.81	(0.39)
Constant	Late	134.1	(82.6)	134.6	(79.1)	7.80	(0.40)
Exponential	Middle	156.0	(64.6)	154.7	(52.0)	7.80	(0.40)
Exponential	Late	109.1	(33.8)	118.1	(47.8)	7.80	(0.40)
Fluctuations							
Constant	Fluctuating	128.4	(72.1)	137.6	(74.0)	7.28	(0.57)
Exponential	Fluctuating	119.8	(55.3)	112.7	(42.3)	7.20	(0.49)

two cases had substantially different population profiles. However, this clearly did not result in a significant effect for the results, as shown. In contrast, when the exponential growth proceeded until 200 generations from the end of the simulation, as in the middle exponential growth case, a clear effect was seen: specifically, a significant reduction in the time to the most recent paternal and maternal common ancestors.

Moving on to the bottleneck results, the only two cases that showed any similarity in their average results were the two middle bottleneck runs, with average profiles for the constant population and exponentially growing population, as shown in Figure 10.6. The time of the bottleneck for these two profiles was between 250 and 300 generations back. For the constant population case, this was 50–100 generations back from the average common ancestor time; although this is a large separation in time the very high variance in the coalescent means that many individual runs will have been affected by the bottleneck, leading to the reduction in the time to the single-parent common

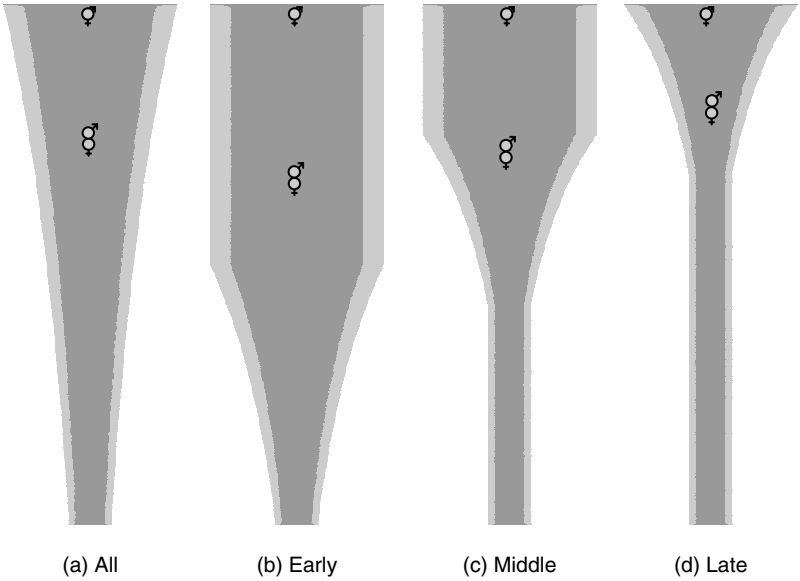


Figure 10.5. Average profiles for the four basic exponentially growing cases, each over 600 generations from a population of 50 to a population of 200: (a) *All*, constant exponential growth; (b) *Early*, exponential growth over the first 300 generations; (c) *Middle*, exponential growth from generation 250 to 450; (d) *Late*, exponential growth over the last 200 generations. Note the influence of the growth period on the common-ancestor generations, as indicated (approximately) by the ♂, ♀ and ⚥ symbols.

ancestors apparent in the average results. In contrast, for the exponentially growing case, the middle bottleneck time was relatively further back, 100–150 generations back from the average common ancestor time, and subsequently had essentially no influence on the result.

More interesting, perhaps, were the later bottleneck results, with the bottleneck occurring at a time when it was almost guaranteed to have a consistent impact on the coalescent tree. Clearly, the two late bottleneck results show a substantial deviation from their respective base cases. When this is related to the number of extant lineages in the respective constant population results, between 100 and 150 generations back, the constant population runs, on average, had from 2.3 to 1.7 of current lineages extant, respectively; in the exponential case, this figure was slightly lower, with from 2.1 to 1.5 current lineages extant.

The bottlenecks had more impact than indicated by the reduction in population size alone, as shown by the plots in Figure 10.7. At the time of the

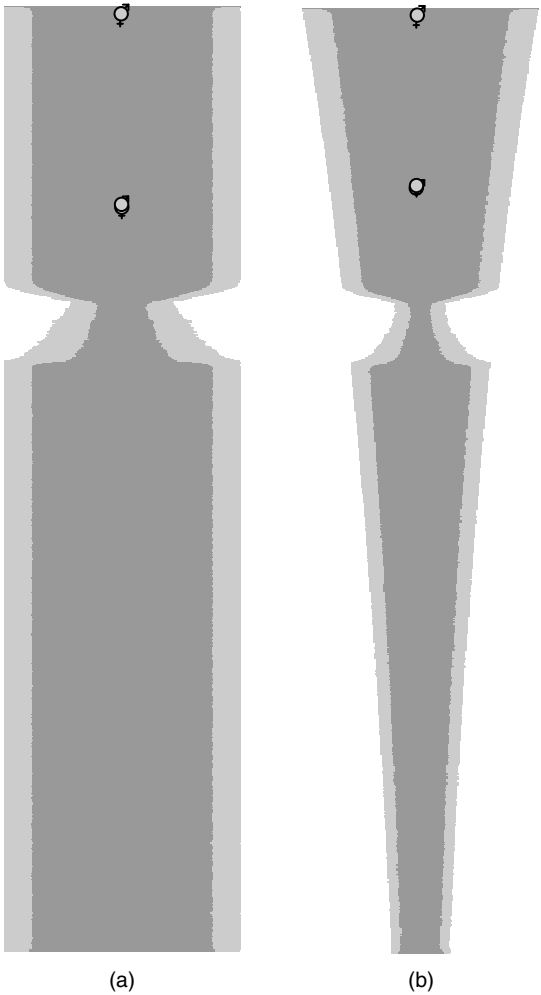


Figure 10.6. Middle bottleneck profiles, averaged over 100 runs, for constant population and exponential growth.

bottleneck, the percentage of individuals ancestral to at least one current individual dropped markedly: from more than 70%, down to around 40% for the duration of the bottleneck, then back to at or near the original value. As can be seen from Figure 10.6, during the bottleneck the population dropped to about 40% of its pre-bottleneck value, so the combined effect of the

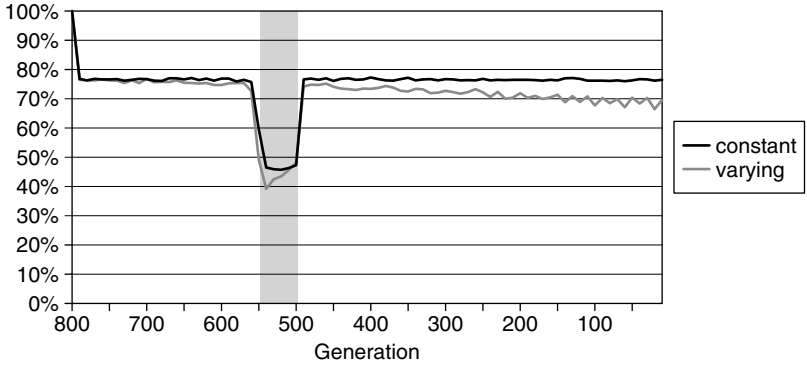


Figure 10.7. Descendant percentages for the two middle bottleneck cases. The shaded area shows the duration of the bottleneck, produced by applying a 60% chance of a major disaster for a period of 50 generations (500–550).

population drop and the much greater degree of lineage extinction illustrates the extent to which a bottleneck affects a genealogy.

This has important implications for the ability of the coalescent tree to give demographic information about the earlier history of a population. For example, from the results in Table 10.6 based on an exponentially growing population, for the late bottleneck case and the late growth case, the generation of the maternal and paternal common ancestors agreed almost exactly, demonstrating how a bottleneck effectively erases the prior demographic information. However, there was a difference in the rate of lineage reduction between these two cases, with the bottleneck case having a sudden drop of lineages as described above and shown in Figure 10.6. Similarly, the late bottleneck run with constant population has both single-parent common ancestors occurring approximately 134 generations back, clearly inside the 100–150 generation range, i.e. the range of the bottleneck, both far more recent than in the purely constant population case.

It is not at all surprising that the biological common-ancestor results were the same for all of the basic profiles and the bottleneck cases,[1] because the rapid mixing of lineages led to a common ancestor many generations before any of these had a significantly different population profile.

However, there was a significant effect in the fluctuating population case for all three kinds of genealogy. The settings that produced the fluctuations were a

[1] ANOVA for the results based on the five basic profiles gave a P value of 0.36, and the lack of difference between the five bottleneck cases is apparent by inspection.

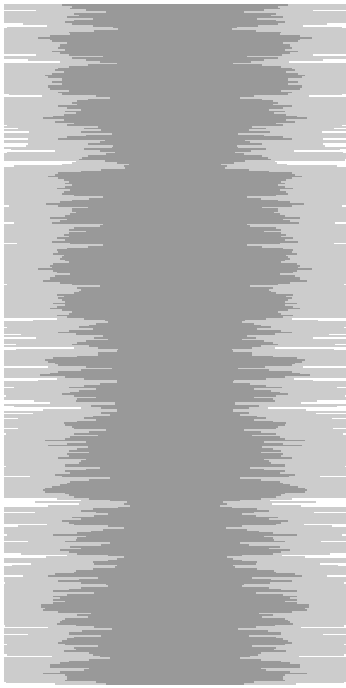


Figure 10.8. A segment of a single (approximately) constant-population monogamy simulation with fluctuation settings as for the average results from Table 10.6, i.e. a 60% chance of a small disaster, a 40% chance of a medium disaster and a 20% chance of a large disaster, applied over all 800 generations.

60% chance of a small disaster, a 40% chance of a medium disaster and a 20% chance of a large disaster, applied over all 800 generations (see Section 9.1 for a description of these parameters). A segment of an example constant population monogamy run with these settings is shown in Figure 10.8. For the 100 simulation average, the average generation size in the constant population fluctuating runs was 193.4, with standard deviation 1.84, and for the survivor percentages, the average number of survivors each generation was 115.3, with a standard deviation of 2.73. However, the averaging masks the true degree of fluctuation. For example, the single runs in Figure 10.8 had very similar average values for both the survivors and the full population size each generation, but the standard deviations were much larger: around 23.5 for the overall population size and 16.4 for the survivors. For the average runs with the fluctuations on top of a basic constant population, the steady state survivor percentage was around 58%, far lower than the nearly 80%

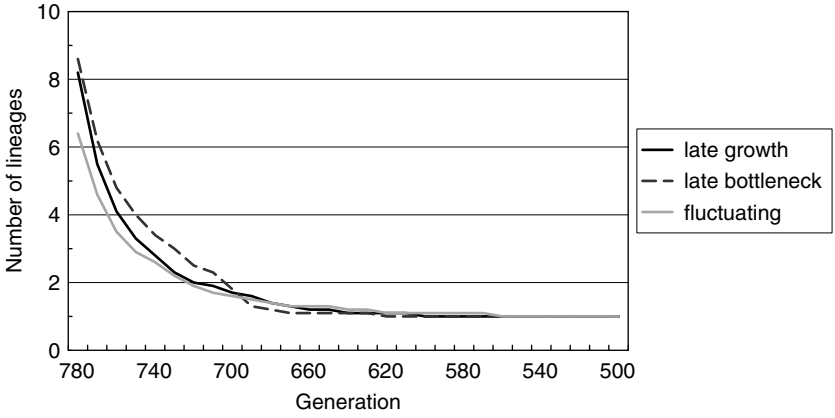


Figure 10.9. Lineages for three exponential growth profiles from Table 10.6: *late growth*, where the period of growth was over the final 200 generations, *late bottleneck*, where there was a bottleneck between generations 650 and 700, and *fluctuating*, where fluctuations are superimposed upon an otherwise exponentially growing population.

figure for the constant population case. These results in particular show the important changes that occur when departing from the simple constant-population coalescent model of Section 8.1.

The average ancestor values for the fluctuating cases in Table 10.6 each match those of one of the other profiles. For the basic constant population case, the fluctuating population results match those of the late bottleneck run. The similarity between the average ancestor values for the late exponential growth population and the exponentially growing population with a late bottleneck has already been mentioned, but these results also match those for the exponential population with fluctuations. These are all qualitatively very different profiles, and the matching in the common ancestor generation clearly demonstrates the subtleties that can make difficult any genealogical or coalescent analysis.

In a real genealogy, it may be possible to use genetic differences to analyse branch lengths and thus differentiate the rate of lineage-merging in the different cases. However, as shown by the similarity of the curves in Figure 10.9, the differences are quite small. In Section 13.5, the possibility of introducing into these simulations the kind of analysis used with the species simulation from the first part of this book is discussed, and this is certainly what is required to model this situation properly.

Table 10.7 shows some similar average values for some alternative mating patterns. Many of the properties seen in the monogamy runs in Table 10.6

Table 10.7. *Varying population results for other mating patterns*

Bottleneck times were between generations 550 and 600 for middle, and generations 675 and 725 for late. The adult sex ratio matches that for the relevant mating groups in all cases.

		Generations back to common ancestor (s_{N-1})					
Basic type	Bottleneck	Paternal		Maternal		Biological	
Polygyny (*MFFF*)							
Constant	None	99.2	(59.6)	273.8	(149.1)	7.16	(0.39)
Constant	Middle	91.5	(41.1)	184.5	(43.3)	7.24	(0.45)
Constant	Late	53.8	(19.1)	94.3	(77.2)	6.95	(0.83)
Exponential	None	84.2	(37.3)	189.8	(77.5)	7.13	(0.42)
Polyandry (*MMMF*)							
Constant	None	250.6	(127.3)	98.0	(57.0)	7.20	(0.40)
Exponential	None[a]	204.6	(92.4)	84.2	(43.5)	7.13	(0.39)

Note: [a] Consanguineous matings are allowed for populations < 50 in the polyandry runs to ensure survival of the population through the early, low population period.

are again apparent. The increased recency of the most recent single-parent common ancestors in the two bottleneck cases was most clearly shown in the late bottleneck case, where the bottleneck occurred around the time of the male common ancestor. The male and female common-ancestor generations were usually well separated for polygynous mating, but in this case the bottleneck forced them both substantially forward in time, the most recent female common ancestor time in particular being nearly 200 generations more recent. In the middle bottleneck case, the bottleneck covered the period between the two ancestors in the purely constant population case, and so had no significant effect on the paternal common ancestor, but the maternal ancestor was approximately 100 generations more recent.

10.2.2 Variation in mating and offspring parameters

Finally, for the single population case, the effect of variation in the mating pattern and fertility parameters during a simulation was examined. Four base cases were employed: monogamy with TFR 4 and reproduction chance 100%, monogamy with TFR 8 and reproduction chance 50%, polygyny with three females per male in each mating group and a 25% male adult sex ratio, and polygyny with the same mating group, but an adult sex ratio of 50% male. In this last case, only one third of all adult males were actually able to mate. On top of each of these base cases, variations in mating pattern, adult sex

Table 10.8. *Average results for simulations with variation from the initial state to the final state as indicated*

The value in the *Change over last* column gives the number of generations at the end of the run over which the change actually occurred. All the runs were for a population of constant size, but, as indicated, the breeding pattern, mating chance, and/or sex ratio varied within the individual runs.

Initial state/ final state	Change over last	Generations back to common ancestor (s_{N-1})					
		Paternal		Maternal		Biological	
Base cases							
M1[a] (TFR/Chance: 4/100%)		193.4	(109.4)	195.2	(101.7)	7.80	(0.40)
M2[a] (TFR/Chance: 8/50%)		81.7	(40.2)	95.0	(55.4)	6.64	(0.48)
P1[a] (50% male)		62.2	(29.1)	166.6	(94.1)	6.46	(0.52)
P2[a] (25% male)		99.2	(59.6)	273.8	(149.1)	7.16	(0.39)
Varying cases							
	800	170.4	(62.0)	196.1	(109.7)	7.87	(0.34)
P1 / M1	400	147.3	(48.4)	190.9	(91.2)	7.88	(0.42)
	200	117.8	(43.1)	172.9	(80.1)	7.88	(0.36)
	800	167.2	(80.3)	187.0	(85.4)	7.85	(0.36)
M2 / M1	400	170.6	(80.4)	168.7	(75.2)	7.84	(0.37)
	200	143.2	(54.5)	135.5	(49.0)	7.89	(0.43)
	800	94.5	(58.8)	101.2	(56.6)	6.55	(0.50)
M1 / M2	400	108.3	(69.9)	119.4	(87.2)	6.63	(0.49)
	200	126.7	(85.4)	134.6	(89.2)	6.71	(0.50)
	800	65.0	(39.0)	168.0	(110.4)	6.45	(0.50)
M1 / P1	400	67.8	(48.6)	164.7	(97.7)	6.49	(0.54)
	200	72.0	(50.8)	165.2	(70.0)	6.46	(0.52)
	800	66.0	(38.7)	170.9	(92.6)	6.58	(0.50)
P2 / P1	400	68.5	(41.2)	186.0	(112.2)	6.57	(0.50)
	200	74.4	(38.9)	238.7	(127.1)	6.54	(0.52)

Note: [a] M, monogamy; P, polygyny.

ratio and reproduction chance were overlaid, with average results as shown in Table 10.8.

The stepwise nature of the mating pattern changes is clearly seen in Figure 10.10, where the survivor percentages indicate the stepping from 1 male/3 female polygyny, to 1 male/2 female polygyny, to monogamy. In contrast, changes in the chance of reproduction are approximately continuous, as seen in the survivor percentages apparent in Figure 10.11.

Each of the varying cases shown in Table 10.8 should be read in conjunction with its start and end states from the base cases, the end case being equivalent to a *Change over last* value of infinity, i.e. no impact on the results from the initial state, and the start case being equivalent to a *Change over last* value of

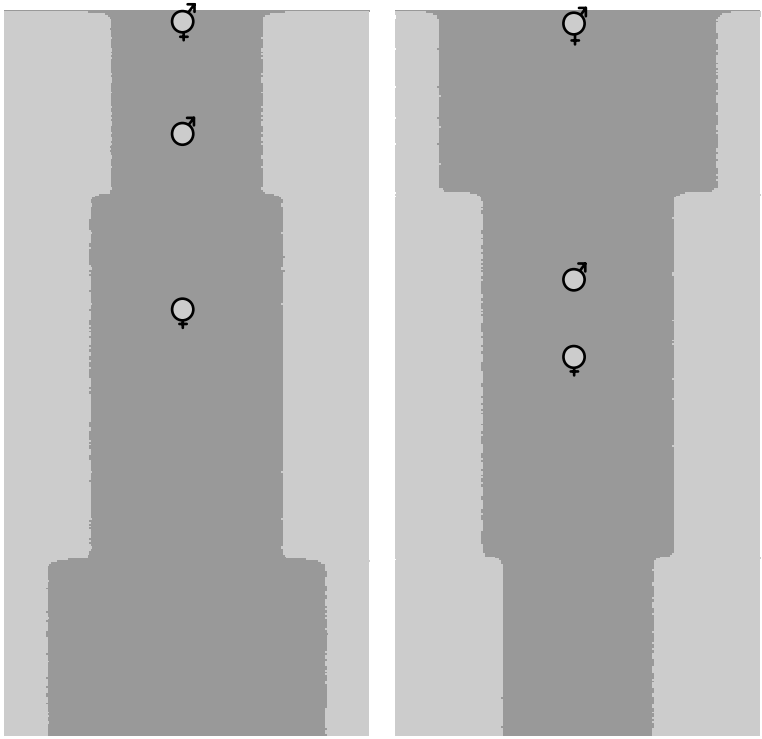


Figure 10.10. Average results for the varying mating pattern simulations within a single population. (a) Variation from polygyny (*MFFF*) to monogamy; (b) the reverse case of variation from monogamy to polygyny. Note the substantial difference in the time to the single-parent common ancestors, especially the paternal common ancestor as indicated by the ♂ symbol.

zero, i.e. no impact on the results from the final state. In this way, the results can be understood as having progressed through five different stages, and the nature of any impact of the changes determined.

The two crucial factors in determining the influence of these changes are the degree of difference between the start and the end states, indicating how much room there is for variation, and also the depth of the genealogy for the end state, indicating how far back in the genealogy it reaches and thus how likely it is to be altered by the earlier state. For example, for the change from polygyny, with a 50% male adult sex ratio, to a monogamous population (see Figure 10.10), there was a much greater difference in the

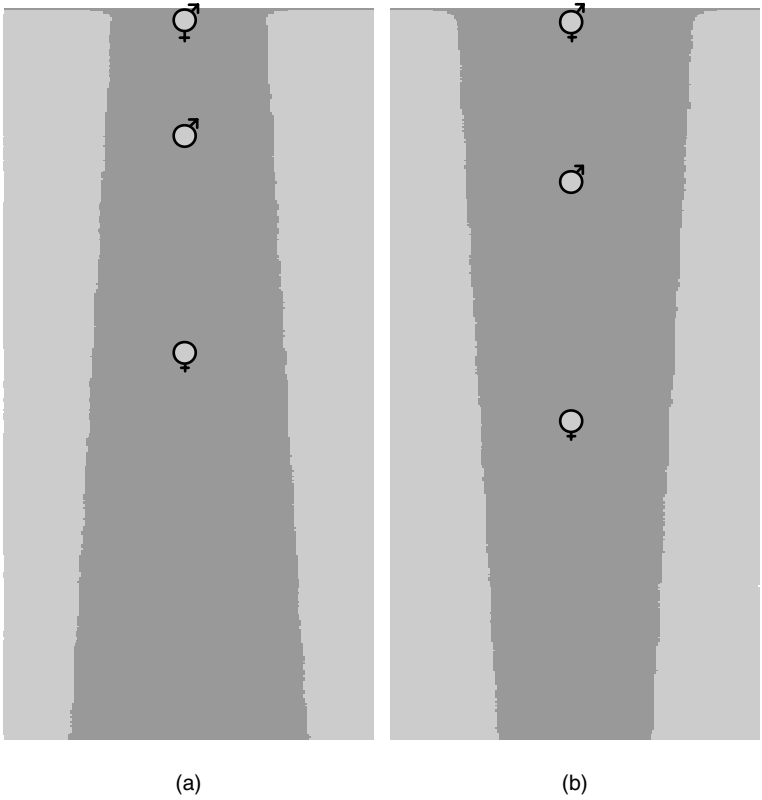


Figure 10.11. Average results for the varying reproduction chance simulations within a single population. (a) Variation in the chance of reproduction from 100% chance per female with TFR 4, to a 50% chance of reproduction per female with TFR 8. (b) The reverse variation. The common ancestors are seen to be further back in (b), because of the greater number of surviving lineages near the end of the simulations in this case.

paternal genealogy than in the maternal one between the start and the end states, and this was reflected in the changes seen for the three cases of a change over 800, 400 and 200 generations. The difference was obvious for the male common ancestor in all cases, but was only apparent for the female common ancestor when the change was over the shortest period. There was no effect on the biological common-ancestor generation, because even over the shortest time span of 200 generations of change the biological lineages of the current population did not *see* any significant demographic change over the $\log_2 N$ time scale of their mixing. In fact, for this very reason, in none

of the cases studied did the biological common ancestor generation show a significant difference from that of the end state.

In the two cases that ended in polygyny with a 50% male adult sex ratio, because of the shallowness of the paternal genealogy (effective population only 50 males) the final state totally dominates the paternal figures and thus obscures the earlier demographic information. This feature was less apparent in the maternal results, but still present, especially in the case of the transition from monogamy with TFR 4 and 100% reproduction chance, where the number of both males and females remains constant at 100 throughout the simulations. When the transition was from polygyny with a 25% male adult sex ratio instead of monogamy, the change in the maternal results was greater because the effective maternal population changed from 150 females down to 100 females over the course of the simulation.

Although the adult sex ratio was constant throughout the simulation in the runs where the change was between the two monogamy states (see Figure 10.11), the effective population changed as a result of the changing reproduction chance, leading to significant differences in the common-ancestor generations in nearly all cases.

11 *Simulating multiple populations*

Adding migrations to the simulation opens up a huge number of further possibilities for investigation. However, rather than attempt an exhaustive coverage, the results in this chapter will instead only cover a handful of particular cases. Less emphasis will be placed on average results than in the previous chapter, since they tend to smooth over and thus obscure many of the more interesting outcomes. Instead, single runs will primarily be discussed, and the impact of migrations examined.

11.1 Sample simulation with regular migrations

In order to provide an illustrative example, the runs presented in this section involved regular *replacement* migrations (see Section 9.1) between three relatively small populations, and ran for only 200 generations. However, some of the additional features of the model, specifically infidelity and non-zero chance of variously sized disasters, were added. The precise settings employed included population size constraints of 40 individuals at generation 10 increasing exponentially to 100 individuals by generation 200, an average fertility rate of 3.5 children and reproduction chance of 90% for the females in the population, and infidelity rates of 20% for males and 10% for females. Constant across all three populations was a 10% chance for a replacement migration involving 10% of the population, and there was also a 10% chance of a small disaster, 5% of a medium disaster and 1% of a large one.

11.1.1 Sample single run

The populations and migrations resulting from one particular run with the above settings are shown in Figure 11.1. With a sample size of 10%, the paternal common ancestor as determined from the sample population matched the true common ancestor for all three populations. However, the timing was quite diverse, with the paternal common ancestor for population *A* only 31 generations back, whereas the paternal common ancestor for the *B* and *C* populations was the same individual, 141 generations back. This individual

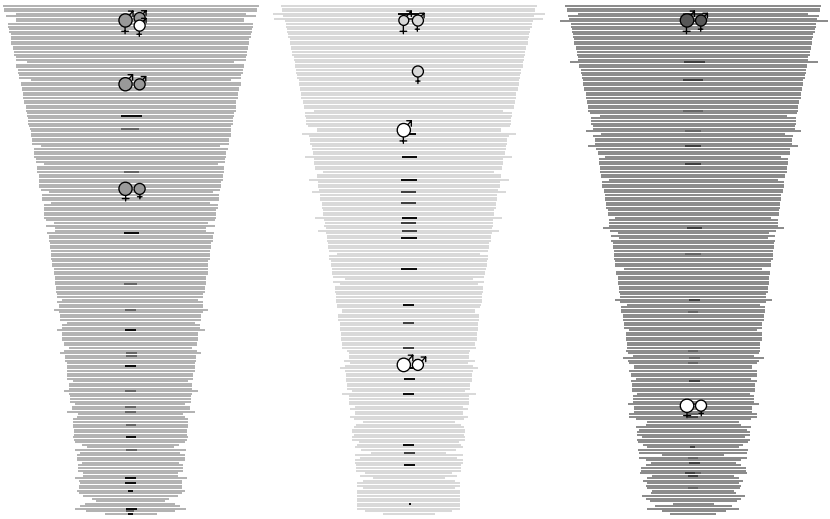
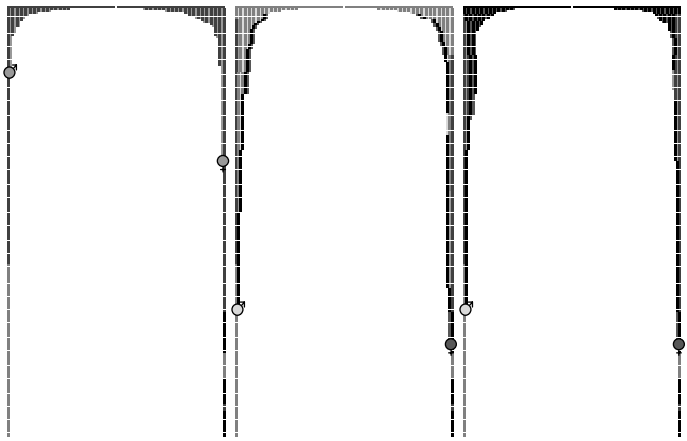


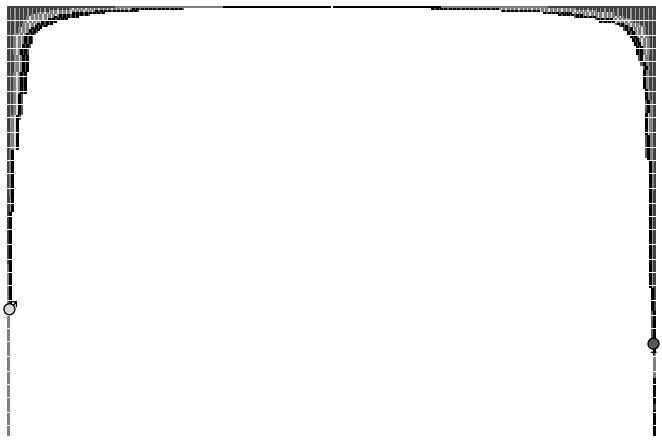
Figure 11.1. Population profiles resulting from the single-run sample simulation. Migrations are indicated by shading them according to their source population, and the most recent common ancestors of the various types are indicated by their respective symbols, the larger ones for the true ancestors and the smaller ones for those determined from the sample population. These symbols are also shaded according to the relevant current population; those in white are the overall ancestors.

was, in fact, the overall paternal common ancestor. (The symbols are therefore superimposed in the figure for these cases.) The situation is similar for the maternal common ancestor: the true *B* and *C* maternal common ancestors are actually the overall ones, but for the *B* sample population the maternal common ancestor is quite recent, only 26 generations back. As indicated by the white ♂ and ♀ symbols in the figure, both the paternal and maternal overall common ancestors were found from the sample population.

The biological common-ancestor analysis produced some interesting results. Similar to what was seen in the single population cases, the most recent biological common ancestor showed little variation between populations or between the true and sample cases. For this run, the most recent biological common ancestor for each population occurred either five or six generations back and was always in the same population as the sample. The most recent biological common ancestor for the overall sample was eight generations back in population *A*. In contrast, the true overall most recent biological common ancestor was actually on population *B* and occurred at generation 151, i.e. 49 generations back. This discrepancy was due to the small number



(a)



(b)

Figure 11.2. Paternal (left) and maternal (right) lineages from each population for the single sample run. True common ancestors are as indicated by the male and female symbols. (a) Individual populations; (b) overall results.

of migrations (simply by chance) between the three populations for the final 50 or so generations, meaning that for this period the genealogies for the individual populations were relatively isolated.

These and further features of the migration history, as it affected common ancestry, are apparent from Figure 11.2. The *B*- and *C*-population common

ancestors matched, and, as discussed above, corresponded to the overall ones. The population-*A* common ancestors occurred much earlier because there was no mixing of lineages from the other populations for much of the later period of the genealogy, and this is seen in the shading of the population-*A* plot in Figure 11.2: purely the population-*A* shade for most of the time shown. In contrast, the lineage colouring for the other two populations clearly shows the impact of migrations in these two cases.

There were 70 migrations altogether over the 200 generations of this simulation, all of them of the *replacement* type. Eight of these migrations were indicated in the full paternal genealogy and 13 in the maternal genealogy, 9 of which were in the generations since the most recent common ancestor. Because of the rapid drop in the number of surviving lineages for the single-parent cases, most of the migrations involved extinct lineages. For the biological genealogy (or pedigree) 65 of the 70 migrations were contained, although only 10 were in the generations since the most recent biological common ancestor.

By 50 generations back in population *B*, and 60 generations back in populations *A* and *C*, the members had fully divided into being either common ancestors of the entire current population, or on extinct lineages. Because of the significant degree of mixing between the three populations, this division occurred for the overall population at about the same time as for populations *A* and *C* because, as is also apparent from the common-ancestor results for these two populations, the mixing within populations is happening largely in parallel with the mixing between populations.

11.1.2 Sample average results

The averages over 100 simulations with these same settings are shown in Table 11.1. A number of interesting features are immediately apparent from these results. For example, the paternal sample common-ancestor generations were much more recent than the true values, especially for the individual populations (because of the smaller size of the male sample). In addition, the dominance of the home population in the ancestor location was clear: very much so for the sample results, and still so, but diluted, for the full population. Because the maternal sample population was twice the size of the paternal sample population, the effect was less significant for the maternal cases. The number of runs where there was no common ancestor was quite high, and highest for the overall common-ancestor case. This is not surprising given that the simulations only ran for 200 generations, but, because of the

Table 11.1. *Results of the average genealogy analyses for paternal and maternal lineages, for both the sample and the full population*

Results are averages over 100 runs. The time back to the most recent common ancestor in each case, along with the standard deviation, plus the distribution of locations are shown. The number of migrations found in the particular genealogies is also shown in the *Migrations found* column, with the figure in brackets being the average number of these migrations occurring after the common-ancestor generation.

Population	Generations back (s_{N-1})	Location				Migrations found
		A	*B*	*C*	none	
Sample population: paternal genealogy						
All	137.5 (39.9)	18%	24%	27%	31%	6.8 (6.3)
A only	73.6 (53.1)	55%	15%	19%	11%	3.5 (2.3)
B only	76.8 (48.2)	16%	53%	18%	13%	3.6 (2.4)
C only	82.1 (55.7)	10%	18%	59%	13%	3.7 (2.5)
Sample population: maternal genealogy						
All	148.8 (34.2)	22%	30%	13%	35%	8.0 (7.7)
A only	107.0 (53.7)	39%	28%	13%	20%	4.3 (3.6)
B only	102.1 (55.7)	17%	55%	13%	15%	4.3 (3.5)
C only	104.3 (55.0)	19%	26%	35%	20%	4.2 (3.4)
Full population: paternal genealogy						
All	145.6 (35.8)	18%	23%	26%	33%	9.0 (8.6)
A only	107.5 (56.2)	39%	18%	21%	22%	4.9 (4.0)
B only	117.1 (47.9)	21%	32%	25%	22%	5.2 (4.4)
C only	107.3 (50.1)	17%	21%	43%	19%	5.2 (4.3)
Full population: maternal genealogy						
All	149.2 (34.4)	22%	30%	13%	35%	9.4 (9.1)
A only	116.8 (51.6)	34%	29%	15%	22%	5.2 (4.6)
B only	119.5 (48.8)	18%	49%	15%	18%	5.2 (4.6)
C only	116.7 (51.8)	19%	29%	29%	23%	5.3 (4.7)

high degree of lineage-mixing and the exponential reduction in population size going back in time, a common ancestor was nevertheless found in a comfortable majority of runs.

The average number of migrations was 58.0. The single-sex genealogies for each of the populations contained four or five of these on average, and the overall genealogies contained approximately nine. The sample-based analysis was only marginally worse in this respect, managing to find three or four of the migrations in each case. These results are consistent with what was seen in the single run described in Section 11.1.1.

The success of the sample analysis in locating the true most recent common ancestor ranged from 49% to 54% for the paternal common ancestor of each

population, up to 86% for the overall paternal common ancestor. For the maternal common ancestor, the corresponding percentages were 77%–79% for the single populations, up to 99% for the overall case. Specifically, 64 of the 65 true common ancestors were found by the sample analysis, the other 35 cases having no overall common ancestor. Consistent with the results from the previous chapter, the biological common-ancestor sample analysis for the individual populations was far less likely to find the true common ancestor, here only 34%–37% for the individual populations, but this increased to 89% for the overall biological common-ancestor case.

The percentage of runs for which the most recent paternal and maternal common ancestors occurred in the same populations ranged from the high 30s up to the low 60s. So in at least one third of runs, and often close to two thirds, these ancestors actually came from different populations!

Figure 11.3 shows the average paternal and maternal lineage distribution for the three populations, and overall. Figure 11.3(a) may be compared with the single-run lineage plot in Figure 11.2. In the single-run case, random effects were more apparent and the lineages were mixed unevenly. In the average case, the lineages in the different populations were much more evenly mixed, because of the smoothing effect of the averaging, thus illustrating the problem of using averaged results to try to understand the behaviour in any particular case. For example, in the average results, the home population dominated its respective genealogies in each case, whereas the single-run results showed how that need not be the case, and indeed, may have significant impact when it is not so (e.g. as in the most recent biological common ancestor result for the single run sample). The average plot indicates a lineage for a particular population even when the average is actually only a fraction of a lineage, and thus, in this example, each case ends in three lines. This simply indicates that the average location of the sole surviving lineage was distributed among the three populations.

11.2 Simulations with restricted migrations

The remaining simulations in this chapter involve a simulated genealogy of 800 generations with three distinct bands of migration: *replacement* migrations across all populations between generation 300 and generation 400, *female-dominated* migrations starting from population *B* between generation 650 and generation 700, and *male-dominated* migrations starting from population *A* between generation 700 and generation 750. This migration pattern is then combined with two different demographic scenarios. Firstly, the

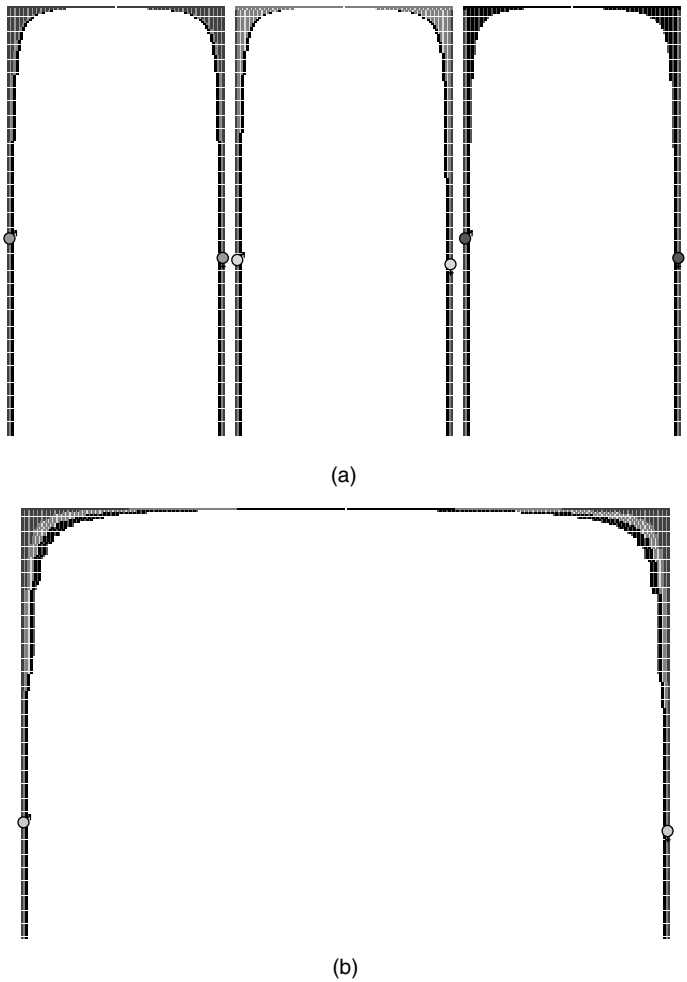


Figure 11.3. Paternal and maternal lineages for the average runs, for both the individual populations (a) and all populations combined (b). As in Figure 11.2, the true common ancestors are as indicated by the male and female symbols.

straightforward case of a constant size, purely monogamous population is simulated and compared against its 100 run average. Then a single run with the same basic situation is studied, but with a bottleneck just before the B-population migrations. All runs are with a constant size of $N = 200$ for each population.

11.2.1 Basic monogamy simulation

The results for a single sample run with a monogamous population, and migration settings of a 10% chance of 25% of the population undergoing a replacement migration between generations 300 and 400 for all three populations, a 25% chance of 10% of the female members of population *B* undergoing migration between generations 650 and 700, and a 25% chance of 10% of the male members of population *A* undergoing migration between generations 700 and 750, are shown in table 11.2.[1] The migrations are shown in Figure 11.4. Interesting points include how the most recent biological common ancestor was very recent for each of the three populations alone, but was several times further back for the overall case. However, examination of Figure 11.4 shows that this overall value was in fact very soon after the first migrations that involved population *A*; thus the final mixing of biological lineages actually occurred very quickly once the population mixing began. There was a very large difference between the overall male and female most recent common-ancestor generations in this simulation. Specifically, the paternal ancestors were all late in the genealogy, and all part of population *A*, because of the migration at this time. In contrast, the most recent maternal common ancestors were each in their home population, as were the most recent biological common ancestors, in the latter case because the short time scale of biological lineage-mixing allowed little chance of participation in even the latest of the migration periods. The overall maternal common-ancestor generation in this simulation relied on the replacement migrations between generations 300 and 350 for the final lineage merging: for the long period between the early and middle migration bands there were two surviving maternal lineages, one in population *A* and one in population *B*. The small numbers of migrations found result from the fact that, even in the deeper single-sex genealogies, only very few of the migrations were seen, because of the speed of the initial coalescences. This was not so for the biological genealogies, because so many lineages were maintained in these cases.

The lineages for the individual populations are shown in Figure 11.5. The lineage plot for population *A* is entirely in the population-*A* shade for the generations shown, and the population-*A* influence on the male lineages is apparent in all 3 cases. Similarly, there is a marked population-*A* influence in the maternal lineages for two of the populations, directly reflecting the

[1] As described in Section 9.1, a small percentage of any migration dominated by one sex is made up of members of the opposite sex.

Table 11.2. *Results of the true common-ancestor analysis for a single constant monogamous population simulation, with the three bands of migration as discussed in the text*

Again, the migration figures in brackets are those after the common ancestor generation. The actual numbers of migrations were 31 replacement, 15 male-dominated, and 10 female-dominated.

Genealogy	Population	Generations back to CA	Location	Migrations found
Paternal	All	84	*A*	5 (3)
	A only	72	*A*	2 (0)
	B only	84	*A*	3 (1)
	C only	84	*A*	4 (2)
Maternal	All	443	*B*	3 (2)
	A only	65	*A*	2 (0)
	B only	159	*B*	1 (0)
	C only	40	*C*	2 (0)
Biological	All	55	*A*	58 (2)
	A only	6	*A*	39 (0)
	B only	7	*B*	44 (0)
	C only	7	*C*	53 (0)

migration settings. All single-population common ancestors occurred prior to the 50 generations of full lineage-mixing earlier in the simulation.[2]

This pattern remained in the 100 simulation average lineage results, shown in Figure 11.6, and the effects the migrations had on the common ancestor locations and times are evident from the results in Table 11.3 and the plot in Figure 11.7. One of the most notable features is the size of the standard deviation in the most recent common-ancestor generation for the single-sex genealogies from the individual populations: often substantially larger than the generation value itself! This is unlike the sample average results discussed in Section 11.1.2, where the migration settings were more even and therefore so were the average common-ancestor results. The overall common-ancestor generations for the paternal and maternal genealogies were very close in time, however, as indicated by the male and female symbols on the left of Figure 11.7. For these cases the migration effects were smoothed over by the averaging because the typical time of the overall common ancestors was significantly further back than the time of the bands of sex-specific migration. The side-effect of population size reduction associated with migrations is also

[2] The lineage plots in particular are much clearer when displayed by the simulation in colour.

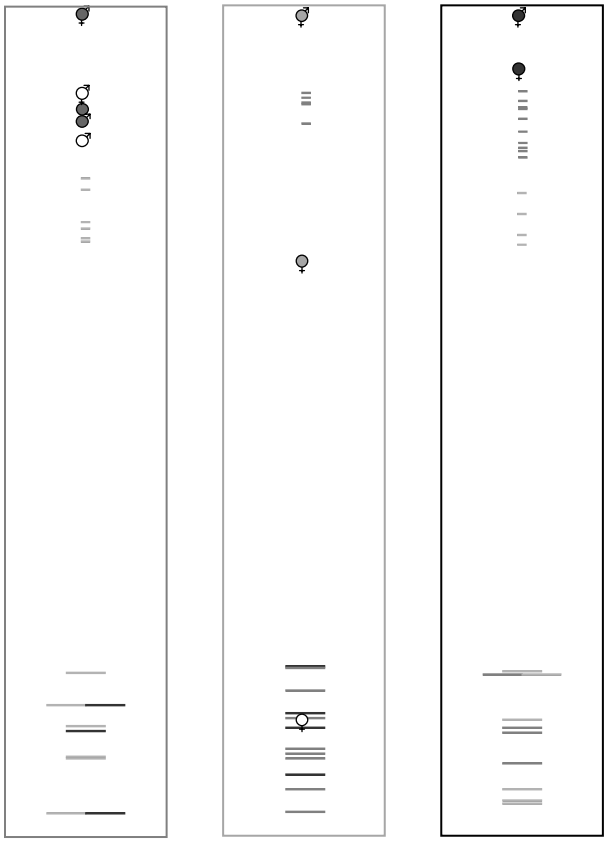


Figure 11.4. A single basic monogamy run with migration, showing the last 550 generations for three populations of constant size (outlined here, rather than with all generations indicated individually as in the actual simulation output) and the migrations between them. The true common-ancestor generations and locations (see Table 11.2) are, as usual, indicated by the superimposed symbols.

visible in Figure 11.7, its appearance marking each of the three migration periods.

Looking first at the single-sex genealogy results only, none of the genealogy analyses for the individual populations failed to find a common ancestor in more than 10% of the runs. In fact, only the full-population paternal case for populations *B* and *C* had no common ancestor for more than 4% of runs, this being the most sensitive case because the male-dominated migrations were the most recent. However, an overall common ancestor was absent much

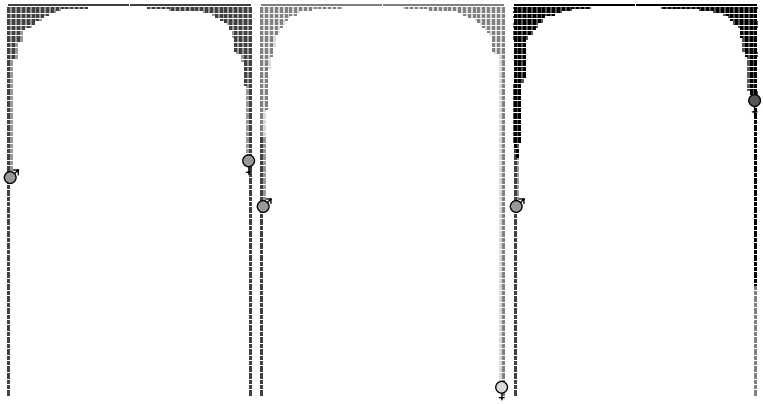


Figure 11.5. Lineages for the single basic monogamy run with migration. Only the final 180 generations are shown, because all common ancestors occurred within this period.

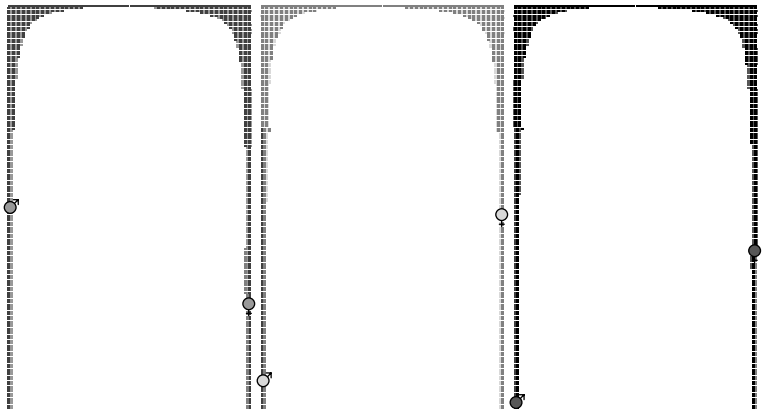


Figure 11.6. Average lineages for the simple monogamy case. As for Figure 11.5, only the final 180 generations are shown.

more frequently, in between 30% and 40% of the simulations. When a male common ancestor was present, it was predominantly in population *A*, and similarly, a female most recent common ancestor was most likely to be found in population *B*, because of the sex-specific migrations. Similar patterns were seen in the sample and full results; the sample actually gave a very accurate picture for the maternal case.

Table 11.3. *Average common-ancestor time and location results for the simple monogamy case, as well as migrations found, for both sample and full populations across all three genealogy types*

Results are averaged over 100 runs.

Population	Generations back (s_{N-1})	Location				Migrations found
		A	*B*	*C*	none	
Sample population, paternal genealogy						
All	390.5 (162.4)	38%	18%	14%	30%	5.1 (4.5)
A only	37.2 (36.7)	100%	0%	0%	0%	2.2 (0.0)
B only	63.7 (116.9)	7%	90%	0%	3%	2.9 (0.6)
C only	70.6 (125.7)	11%	0%	86%	3%	3.0 (0.5)
Sample population, maternal genealogy						
All	422.1 (135.1)	13%	31%	18%	38%	5.7 (4.9)
A only	97.1 (117.9)	88%	9%	1%	2%	3.2 (0.5)
B only	63.9 (45.6)	0%	100%	0%	0%	2.6 (0.0)
C only	78.5 (86.8)	0%	6%	93%	1%	3.3 (0.3)
Sample population, biological genealogy						
All	63.78 (9.52)	99%	1%	0%	0%	56.5 (3.6)
A only	4.76 (0.62)	100%	0%	0%	0%	36.5 (0.0)
B only	4.93 (0.64)	0%	100%	0%	0%	42.7 (0.0)
C only	4.93 (0.67)	0%	0%	100%	0%	50.2 (0.0)
Full population, paternal genealogy						
All	423.6 (134.4)	33%	20%	15%	32%	5.9 (5.3)
A only	83.3 (47.8)	100%	0%	0%	0%	2.2 (0.0)
B only	155.3 (164.1)	21%	69%	3%	7%	3.3 (1.5)
C only	164.3 (171.6)	27%	6%	58%	9%	3.6 (1.6)
Full population, maternal genealogy						
All	427.0 (130.1)	13%	30%	18%	39%	5.8 (5.0)
A only	124.0 (124.4)	85%	9%	2%	4%	3.3 (0.7)
B only	86.9 (45.2)	0%	100%	0%	0%	2.6 (0.0)
C only	101.0 (103.4)	1%	7%	91%	1%	3.3 (0.4)
Full population, biological genealogy						
All	63.78 (9.52)	99%	1%	0%	0%	56.5 (3.6)
A only	6.54 (0.52)	100%	0%	0%	0%	36.5 (0.0)
B only	6.48 (0.50)	0%	100%	0%	0%	42.7 (0.0)
C only	6.59 (0.51)	0%	0%	100%	0%	50.2 (0.0)

Matches between the most recent common ancestors, as determined from the sample population and the true ancestors, again tell an interesting story. For the paternal ancestors, there was a match in 31%, 26% and 31% of simulations for the individual population common ancestors, respectively; for the maternal ancestors this increased to 62%, 57% and 64%, and, as before, the biological

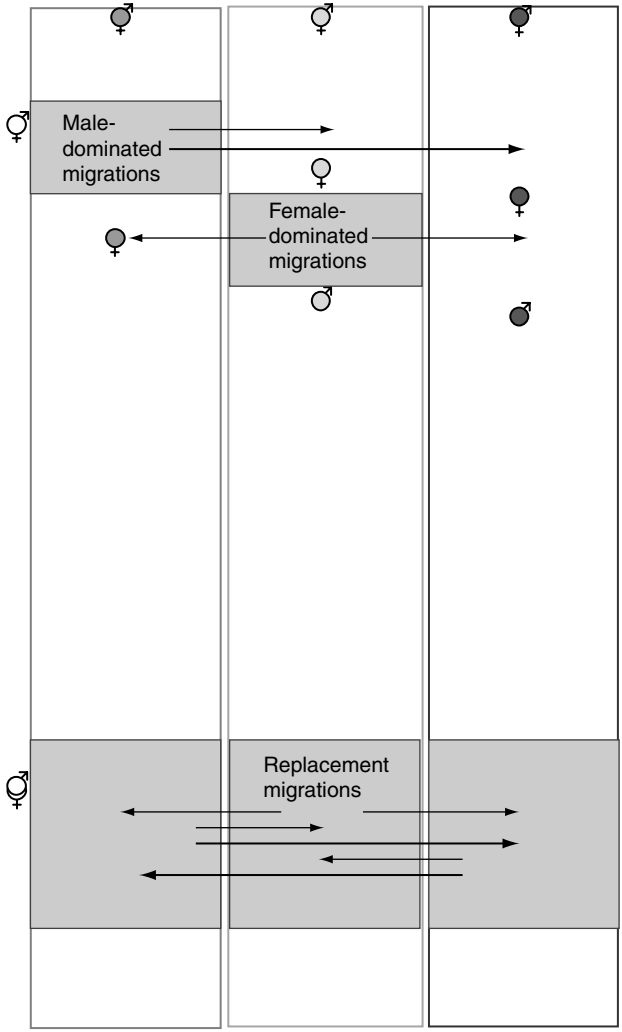


Figure 11.7. Average profiles and common ancestors for the basic monogamy simulations with migration. As for Figure 11.4, only outlines of the last 550 generations of the populations are shown; the periods and types of migration are as indicated.

common-ancestor matching was by far the least successful, with success percentages of only 3%, 9% and 8%. However, for the overall common ancestor, the figures rose to 84% for a paternal most recent common-ancestor match, 97% for a maternal most recent common-ancestor match, and

precisely 100% for a biological match: the true overall biological ancestor was found from the sample analysis in every one of the 100 simulations! This improvement in success was because the biological lineages from the individual populations were already thoroughly mixed at the time when the migrations began, so although the individual population samples found a common ancestor slightly early in each case, by the time the populations began to mix, the lineages traced back from the sample population and the full lineages were essentially identical.

11.2.2 Basic monogamy with a bottleneck

Once again considering a single run with the previous settings, but with the addition of a relatively brief, 30 generation, bottleneck on population *B* from generations 600 to 630, resulted in the population profiles shown in Figure 11.8. The reduction in the number of lineages in population *B* at the time of the bottleneck was very significant for the simulation because of the subsequent female radiation from that population, and thus the lineages participating in the radiation were more closely related than would have been the case had the constant population been maintained. As individuals from population *B* migrated and mixed with the other populations, this resulted in a relatively recent overall common ancestor for maternal lineage as seen, occurring just after the bottleneck. There was no overall male common ancestor for this simulation, owing to insufficient mixing of male lineages in the two bands of migration that included a significant migrating male component.

Several possible scenarios such as this one can easily be studied by using the simulation, each illustrating important aspects of genealogies under various demographic constraints. For example, many more runs could be made to study the interplay between different kinds of migration and different mating patterns and sex ratio, etc. In addition, all the results presented in this chapter have been for a population of constant size. Further runs could study different kinds of population size change, as was done in the single-population case in the previous chapter, and with the multiple-population model these changes could be different for each population. However, even though all these questions are interesting in their own right, in order to get maximum value from the simulations the addition of genetics to the genealogy is essential. So, for example, instead of simply studying the mixing of lineages when members of a declining population migrate and merge into a population of constant size, the situation can be made even more interesting by bestowing some degree of selective advantage on the migrants.

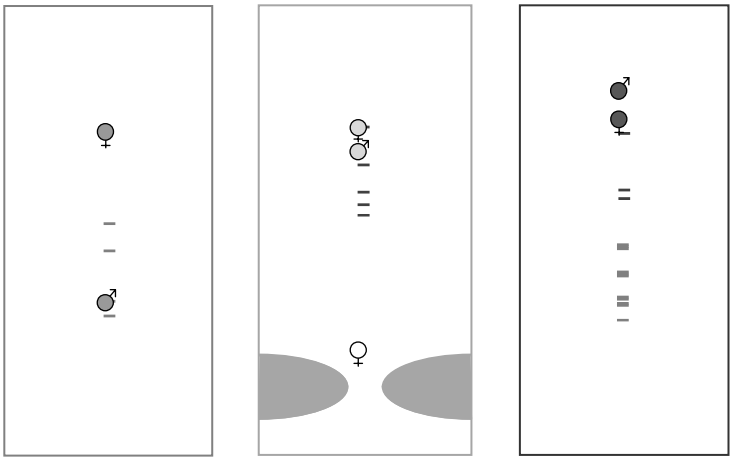


Figure 11.8. The final 230 or so generations for a single constant-population monogamy simulation, with a bottleneck on population *B* just prior to the start of the female-dominated migrations from that population (as indicated by the shaded regions). An overall male most recent common ancestor was not found; and the biological ancestors are not shown, because they are unaffected by a bottleneck so far back in time.

With this in mind, the next chapter will present the development of the two genetics models that are included with the *Genie* simulation, and illustrate their use through simulation of a scenario that extends the restricted and varying migration results of Section 11.2 by including both haploid and diploid neutral genetics, selective advantage and polygyny.

12 *Adding genetics to the genealogy*

Although studying genealogies is interesting in its own right, it is clear that the addition of genetics is fundamentally important for applying the simulation to problems relating to human origins. To this end, two independent genetics models have recently been added to the *Genie* application and are the subject of ongoing development. The reader is referred to the website for updated details and results related to the genetics models.

The first of these involves neutral genes only, and therefore is simply superimposed on the simulated genealogies without affecting in any way the survival or reproductive success of the carriers. Six loci in all are modelled: two on each of the Y chromosome, an autosome, and mitochondrial DNA. The second genetics model involves two alleles at a single locus and covers selection, dominance and mutation between the two alleles.

Each of these models is described in detail below, but first the modelling of genetics in the context of coalescent theory is presented, along with methods for estimating the mutation rate and time to the most recent common ancestor.

Finally, results of simulations that combine restricted and sex-specific migrations with both neutral and non-neutral genetics are shown and discussed.

12.1 Modelling genetics with coalescent theory

As for the earlier results, the theory relating to the simulated genealogies is best described by using coalescent theory, even when genetics is included. The theory presented below is again for a single, randomly mating population of constant size, but has extensions similar to those of the basic genealogy theory presented in Section 8.1.

Neutral mutations are modelled as a Poisson process with constant rate $\theta/2$, where

$$\theta = \begin{cases} 4N_e\mu & \text{for nuclear genes} \\ 2N_e\mu & \text{for mtDNA or Y-chromosome genes,} \end{cases} \tag{12.1}$$

N_e being the appropriate effective population and μ the mutation rate per gene per generation. This process occurs independently on each branch of the genealogy and so, if we assume that the mutation rate is sufficiently low that each mutation occurs at a new site, the number of segregating sites in a sample of size n is equal to the number of mutations, and has mean

$$S_n = \frac{\theta}{2} L_n = \frac{\theta}{2} \sum_{j=2}^{n} jT(j) \tag{12.2}$$

for tree length L_n, where $T(j)$ is the number of generations with j distinct ancestors, given by Equation 8.5. This approach is known as the infinite sites approximation. Closely related is the infinite alleles approximation, where each mutation is assumed to introduce a new allele into the population (Hartl and Clark, 1997).

12.1.1 Estimating the mutation rate

The above discussion describes a simple mutation model, but the more fundamental problem is the backward-looking one: given a sample, what can be determined of the population parameters? Equation 12.1 relates the population size and mutation rate; given a sample of size n, there are various estimators of the parameter θ; see, for example, the review in Donnelly and Tavaré (1995).

Related to the discussion in Section 12.1 is Watterson's estimator (Watterson, 1975) based on the number of segregating sites:

$$\theta_W = S_n \left(\sum_{k=1}^{n-1} \frac{1}{k} \right)^{-1} \tag{12.3}$$

with variance

$$\mathrm{Var}(\theta_W) = \left(\theta \sum_{k=1}^{n-1} \frac{1}{k} + \theta^2 \sum_{k=1}^{n-1} \frac{1}{k^2} \right) \left(\sum_{k=1}^{n-1} \frac{1}{k} \right)^{-2}. \tag{12.4}$$

Alternatively, according to the infinite alleles model, the estimate θ_E can be derived from the Ewens sampling formula (Ewens, 1972) and is the solution of

$$K_n = \sum_{k=0}^{n-1} \frac{\theta}{\theta + k}, \tag{12.5}$$

where K_n is the number of distinct alleles in the sample. This estimator has variance

$$\mathrm{Var}(\theta_E) = \theta \left(\sum_{k=1}^{n-1} \frac{k}{(\theta+k)^2} \right)^{-1}. \tag{12.6}$$

In addition, Tajima (1983) proposed the following estimator based on pairwise differences:

$$\theta_T = \frac{2}{n(n-1)} \sum_{i<j} \pi_{ij}, \tag{12.7}$$

where π_{ij} is the number of sites at which sequences i and j differ. This estimator has variance

$$\mathrm{Var}(\theta_T) = \frac{n+1}{3(n-1)}\theta + \frac{2(n^2+n+3)}{9n(n-1)}\theta^2. \tag{12.8}$$

The estimators θ_W and θ_E have variance that approaches zero as $n \to \infty$, and as such are consistent, although the zero limit is approached at a rate of $1/\ln n$. The estimator θ_T has a non-zero limit as sample size increases, and is therefore not a consistent estimator. As discussed in Donnelly and Tavaré (1995), this is because the sampled genes are not independent but share a genealogy. Fu (1994) studied the problem of minimising variance in mutation-rate estimators and proposed an estimator that makes full use of phylogenetic information in a sample of DNA sequences from a population. This estimate has variance that is only slightly larger than the minimum possible and is substantially smaller than that of the estimators described above.

A related measure is Tajima's D (Tajima, 1989), which, given a sample, tests the hypothesis of selective neutrality. It is based on the observation that, because θ_T and θ_W are essentially measuring the same thing, in the case of selective neutrality they should be equal, and is defined as

$$D = \frac{\theta_T - \theta_W}{\mathrm{Std}(\theta_T - \theta_W)}, \tag{12.9}$$

where $\mathrm{Std}(\theta_T - \theta_W)$ is the standard deviation of the difference between the two estimates. If $D \neq 0$ with statistical significance, then departure from the underlying assumptions is indicated. For example, an excess of

rare alleles will result in a negative value for D, whereas balancing selection will increase nucleotide diversity and result in a positive value for D. However, this measure is also sensitive to population history and structure, and thus non-zero D values do not unambiguously indicate a departure from selective neutrality. Alternative tests are discussed by Neilsen (2001).

12.1.2 Estimating the TMRCA

Before considering genetic data, the genealogy of a Wright–Fisher population is as described in Section 8.1, with mean time to the most recent common ancestor $T_{\mathrm{MRCA}} \approx 2N$ for a population of constant size N, and distribution a sum of exponential distributions for each period of j lineages as described by Equation 8.4. However, by following a Bayesian approach where T_{MRCA} is the underlying variable with prior distribution given by the coalescent as above, and posterior distribution determined in the light of the measured genetic data, this estimate may be revised. Examples of this approach include the acceptance–rejection algorithm of Tavaré *et al.* (1997), or model parameter estimation based on the maximum likelihood of the data (Griffiths and Tavaré, 1994, 1996). In general, the solution is both theoretically and computationally complex. However, an analytical solution is available in the following two limits.

Tajima (1983) found exact solutions for the mean and variance of T given a sample of size 2 and k segregating sites:

$$E(T_{\mathrm{MRCA}}|S_n = k) = \frac{\theta(1+k)}{2\mu(1+\theta)}$$

and

$$\mathrm{Var}(T_{\mathrm{MRCA}}|S_n = k) = \frac{\theta^2(1+k)}{4\mu^2(1+\theta)^2}.$$

The other case that permits exact solution is for samples of any size that display no variation (Fu and Li, 1996; Tavaré *et al.*, 1997). Under such circumstances

$$E(T_{\mathrm{MRCA}}|S_n = 0) = 2N\sum_{i=2}^{n}\frac{1}{i(i-1+\theta)}$$

and

$$\mathrm{Var}(T_{\mathrm{MRCA}}|S_n = 0) = 4N^2 \sum_{i=2}^{n} \frac{1}{i^2(i-1+\theta)^2}.$$

Tang *et al.* (2002) present an alternative estimation formula, based on the molecular clock hypothesis and not requiring any population structure assumptions.

12.1.3 Recombination

Recombination is an important source of genetic variability and occurs when gametes in offspring do not match those of their parents, but result from a rearrangement of genetic material. For example, consider the two-locus parental genotype AB/ab. Then gametes AB and ab are directly inherited and match the parental gametes, whereas gametes Ab and aB result from recombination. Analogous to Equation 12.1, recombination is described by the parameter

$$R = 4N_e r, \tag{12.10}$$

where r is the rate of recombination per generation. A large value of R implies independent genealogies for each locus, whereas a value of zero implies precisely the same genealogy at each locus.

The property that when there is recombination there can be more than one evolutionary history underlying a single sample is illustrated by Figure 12.1. The figure shows a two-locus ancestral recombination graph and the associated marginal coalescent trees for each locus. Going backwards, recombination events result in an increase in the number of lineages, and coalescent events in a decrease. The figure is drawn so that, at a branch point, the locus 1 branch is to the left and the locus 2 branch to the right.

Extensions to coalescent theory to handle recombination were first presented by Hudson (1983) and are reviewed by Hudson (1990). Given j lineages and an 'event', the probability that the event is a coalescence is given by

$$P(j \to j-1) = \frac{j-1}{j-1+R}.$$

with $R = 4N_e r$ as above. Similarly, the probability it is a recombination is given by

$$P(j \to j+1) = \frac{R}{j-1+R}.$$

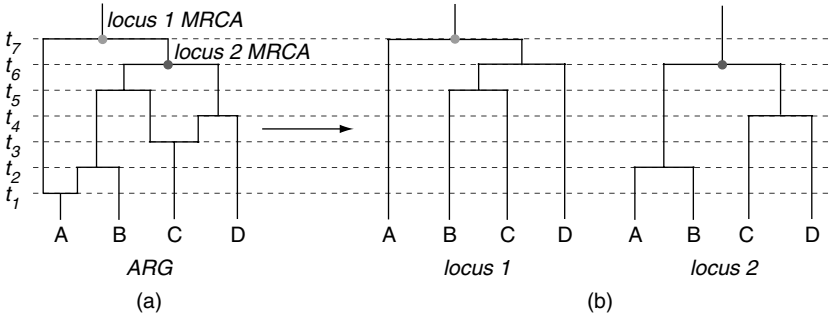


Figure 12.1. Ancestral recombination graph (*ARG*) for a two-locus sample, and the associated marginal coalescent trees for each locus. The first event, at time t_1, is a recombination involving sample A, and the locus 2 sample A allele clusters with sample B at time t_2, whereas the locus 1 sample A allele remains distinct all the way back to time t_7 at the common ancestor of the whole sample. The other recombination involves sample C at time t_3, and the locus 1 sample C allele clusters with sample B at time t_5, whereas the locus 2 sample C allele clusters with sample D at time t_4. Each locus has a different most recent common ancestor, as shown by the different marginal coalescent trees.

The distribution of the time to any event is exponential, with expected value

$$E(T(j)) = \frac{4N}{Rj + j(j-1)}. \tag{12.11}$$

This is a result of the two independent rates: $Rj/4N$ for lineage birth, and $j(j-1)/4N$ for lineage death. Because the birth rate is linear and the death rate quadratic, the value $j = 1$, and thus a common ancestor, is reached with certainty. Equation 12.11 is the analogue of Equation 8.5 for a diploid population with recombination.

In human populations, recombination and mutation rates are estimated to be of the same order of magnitude (Nordborg, 2001a). Posada *et al.* (2002) discuss methods for detecting recombination and measuring the rate, Hey and Wakeley (1997) present a coalescent theory method for estimating R, and Griffiths (1999) studies the time to the ancestor along sequences with recombination.

Dealing with recombination clearly adds a great deal of complexity; indeed, most recombination events are undetectable. However, when there is a high degree of recombination in some region of DNA, the unlinked loci actually provide many independent samples of the genealogy, thus averaging out some of the evolutionary stochasticity. Consequently there is a reduction in the variance of most test statistics, such as the theta estimators

presented in Section 12.1.1, in the presence of recombination (Hudson, 1990; Nordborg, 2001a).

12.1.4 Linkage disequilibrium

An important concept in population genetics is linkage disequilibrium (Hartl and Clark, 1997; Nordborg and Tavaré, 2002), which provides a measure of the non-random association between alleles at different loci. Differences from purely random association are generated, by selection or migration for example, and decay slowly over time through the action of recombination (Zondervan and Cardon, 2004).

Linkage disequilibrium can be used to study population histories, both size (Hayes *et al.*, 2003), because the rate of recombination is related to the effective population through Equation 12.10, and structure (Falush *et al.,* 2003), because mixing of populations with different gene frequencies will produce linkage disequilibrium that will persist for some generations.

Linkage disequilibrium helps to reveal the location and relationships of the genes along the chromosomes and is therefore particularly important for attempts to refine the location of disease genes by detecting differences in marker loci between affected and unaffected individuals (Jorde, 1995; Nielsen *et al.*, 1998; Nordborg and Tavaré 2002). This is especially significant for rare disease alleles where sufficient fine-scale pedigree and recombination data are not available.

However, linkage disequilibrium is an awkward concept to define, presenting difficulties of interpretation even in simple limits (Lewontin, 1988), and may be quantified in several ways (Hedrick, 1987). One commonly used definition arises by considering two loci with alleles A_i and B_j and respective frequencies p_i and q_j. Then a measure of the pairwise linkage disequilibrium between alleles A_i and B_j may be defined as

$$D_{ij} = h_{ij} - p_i q_j, \tag{12.12}$$

where h_{ij} denotes the population proportion of haplotype A_iB_j. Obviously $p_i = \Sigma_i h_{ij}$ and $q_i = \Sigma_j h_{ij}$.

To enable comparisons across loci and populations, this measure may be normalised by dividing by

$$D_{\max} = \begin{cases} \min(p_i q_j, (1-p_i)(1-q_j)) & \text{if } D_{ij} < 0 \\ \min(p_i(1-q_j), (1-p_i)q_j) & \text{if } D_{ij} > 0, \end{cases} \tag{12.13}$$

to produce the measure

$$D'_{ij} = D_{ij}/D_{\max}, \tag{12.14}$$

which ranges between -1 and $+1$.

If there are only two alleles at each locus, D'_{ij} is unique. Otherwise, the sum

$$D' = \sum_{ij} p_i q_j D'_{ij}, \tag{12.15}$$

which will range between 0 and 1, may be used (Hedrick, 1987; Zapata *et al.*, 2001).

Extending this measure to handle more than two loci adds substantial further complexity, as discussed by Ayres and Balding (2001).

12.1.5 Migration and selection

Two other extensions to coalescent theory that need mention are migration and selection. Although both are of fundamental importance, they will only be covered very briefly here.

The most commonly employed migration model is the island model, where there is assumed to be a Wright–Fisher population on each of a (possibly infinite) number of 'islands', and a factor m_{ij} governs the degree of migration from island i to j each generation. When incorporated into coalescent theory, this leads to the *structured coalescent* (Hudson, 1990, 1998). The basic approach is similar to that employed in Section 12.1.3 above, where the coalescent tree consists of a sequence of *events* and on going back in time each event is either a coalescence, with rate as for the standard coalescent, or a migration, with rate determined by m_{ij}. For an application, see, for example, Beerli and Felsenstein (1999).

Incorporating selection into a coalescent framework can proceed similarly, with mutation between 'allelic classes' analogous to migration between 'islands' (Nordborg, 2001a). Alternatively, Neuhauser and Krone (1997) have introduced the concept of the *ancestral selection graph*, that is a graph showing all potential ancestral paths, with branch points corresponding to selection acting on alleles of differing fitness. Thus the ancestral selection graph is constructed going backwards in time, containing coalescences as normal and branches indicating potential descent associated with selection. Mutations are then added going forward in time, and the graph is 'pruned' based on the fitness of the potential branches, leading to the embedded true genealogy.

12.2 Genetics models in the simulation

As mentioned earlier, there are two distinct genetics models included in the simulation. The first of these is a neutral model and superimposes independent and variable mutation for six loci (two on each of the Y chromosome, an autosome and mitochondrial DNA) and variable recombination for the autosome loci on the simulated genealogy. Independently, two alleles at a single locus (which may optionally be X-linked) are modelled, with parameters controlling the degree of selective advantage, dominance and mutation between the alleles. Because of the role of selection in this case, the model is not independent of the genealogy, and the survival of individuals is directly affected by their genotype.

12.2.1 Neutral model

Each locus in the neutral genetics model has an independent mutation rate, and this rate may be constant or vary linearly over a range of generations. All genes are currently implemented as bit strings of length 16, and mutation acts by toggling an individual bit. Recombination is implemented for the autosome and, as for the mutation rate, may be constant or vary linearly over a range of generations. Only recombination between genes is modelled at this stage: there is no intragenic recombination.

The initial genetic diversity is controlled by specifying the number of alleles at each locus for each population, and whether the three populations share these initial alleles or whether they are independent. The set of initial alleles is constructed by first randomly toggling one of the 16 bits and subsequently toggling initial additional bits until a unique allele has been constructed. There is also some modelling of gene expression through 'characters' that depend on the genotype in a deterministic but non-trivial way.

The textual output relating to the genetics simulation is quite detailed. In particular, it lists the complete genotype and characters for each member of the final generation, because this is the primary data for any backward-looking analysis of the population.

Because of the amount of historical genetic data generated, it is necessarily provided in a more summarised form. For each generation simulated, the number of distinct alleles at each locus is reported, along with the average pairwise difference and number of segregating sites. A more detailed look at the historical genetic data is provided by snapshots every ten generations, listing the alleles in each population at that generation, giving their frequency and, for the autosomes, the number of homozygotes. For the character data,

the number of distinct characters and their average pairwise difference is reported for each generation.

The graphical display allows any allele to be selected, and its corresponding distribution is displayed as shown in Figure 9.1c (p. 156).

The mutation-rate estimators presented in Section 12.1.1 are calculated for each of the six loci simulated, based on both a sample from, and all members of, each population separately and for the overall population. The linkage disequilibrium each generation is also calculated for each population and overall.

In addition to the purely genealogical common-ancestry information discussed in Section 9.3, and used substantially in analysing the simulation results presented in the previous two chapters, the time and location of the most recent common ancestor for each autosome locus (because they may be different when recombination is included) is calculated and included in the text report.

12.2.2 Genotype model

The genetics model that simulates selection is distinct from the neutral model described in the previous section, and models two alleles at a single locus. The two alleles are labelled A and a, and the locus may be X-linked. In order to simulate selection, one of these alleles must be identified as favoured; for the purposes of the following discussion, call this allele F and the other allele X.

Specification of the model then involves setting the following parameters. The *selection coefficient*, s, ranges from 0 to 1 and specifies the degree of advantage conferred by the favoured allele F. The *degree of dominance*, h, also ranges from 0 to 1, with $h = 0$ implying complete dominance of the favoured allele and thus the fitness of genotype FX equivalent to that of genotype FF, $h = 1$ implying complete dominance of the non-favoured allele and thus genotype FX equivalent to genotype XX, and intermediate values $0 < h < 1$ implying incomplete dominance, with dominance of the favoured allele decreasing as h increases. The *heterozygote fitness factor* ranges between -1 and 1, with $z > 0$ implying overdominance and genotype FX having the greatest fitness, and $z < 0$ implying underdominance, where genotype FX is of reduced fitness, possibly even more so than a homozygote in the non-favoured allele, genotype XX. For any simulation, only one of h and z may be non-zero.

The fitnesses are then determined as shown in Table 12.1, and the implementation in the simulation is to treat the fitnesses as indicating the relative probability for individuals of each genotype of surviving to mating.

Table 12.1. *The fitnesses for each genotype give the relative probabilities of survival to mating*

Note that the maximum value in each row is 1, and if $1-s-z<0$, the fitness is treated as 0.

	FF	FX	XX
$z>0$, $h=0$	$1-z$	1	$1-s-z$
$z=0$	1	$1-hs$	$1-s$
$z<0$, $h=0$	1	$1+z$	$1-s$

By default, the generation-1 population is entirely of genotype aa. The introduction of allele *A* into the simulation is by optionally specifying a generation for each population where a certain percentage of homozygous AA individuals, and a percentage of Aa individuals, will be introduced. In general, the generation chosen will be very early in the simulation. However, if allele A has no selective advantage, it can be interesting to introduce it to one of the populations at an arbitrary time, and it will subsequently act as a marker.

User-specifiable mutation probabilities control mutation between alleles A and a, with forward and backward rates specifiable independently. These rates may be either constant or linearly varying over a range of generations.

The output includes the distribution by generation for each of the genotypes AA, Aa and aa for each population and overall, in both textual and graphical formats.

12.3 Sex-specific migrations and selection

Continuing on from the simulations presented in Section 11.2, but changing the mating pattern from monogamy to polygyny (with three females for every male and a 50 : 50 adult sex ratio), the results in this section illustrate the use of the simulation to study the effects of sex-specific migrations when the mating pattern is not monogamy and when selective advantage is acting in the population. Changes to the mating pattern affect the paternal and maternal genealogies differently, as discussed in Section 10.2.2; in combination with migration and selection, a variety of different interactions are possible.

The migration settings used are narrower than those used in Section 11.2, with settings of a 10% chance of 25% of the population undergoing a replacement migration between generations 300 and 350 for all three populations, a 25% chance of 25% of the female members of population-*B* undergoing migration between generations 675 and 700, and a 25% chance of 25% of the

male members of population *A* undergoing migration between generations 725 and 750.

Both genetics models are utilised in these simulations. The neutral model starts with five distinct alleles at each locus and a mutation rate of one mutation per gene per individual per 1000 generations everywhere, except for the mtDNA at locus 1 and the Y chromosome at locus 2, which are set to mutate five times more slowly. The recombination rate is similarly set to once per individual per 1000 generations. The genotype model is set to include selection effects, with allele A favoured, with selection coefficient $s = 0.2$, and semi-dominant, $h = 0.5$ (see Section 12.2.2). Neither allele under this model is mutating for these runs. In addition, the average fertility rate was increased for these runs to ensure stable population generation given the introduction of a deleterious allele.

With a mating pattern of polygyny, the most recent maternal common ancestor for a population will on average be further back in time than the most recent paternal common ancestor. However, for the single simulation shown in Figure 12.2, the population-*B* paternal common ancestor is almost 500 generations further back in time than the corresponding maternal ancestor. This is a result of the last two surviving paternal lineages on population *B* being isolated from each other, with one surviving in population *B* and the other in population *A* from generation 744 to generation 309. Owing to the replacement migrations between generations 300 and 350, these two lineages are brought together in population *B* for the nine generations between generation 308 and generation 300, and then in population *A* between generations 299 and 243, before finally coalescing at generation 242.

This further illustrates the extent to which the addition of migrations can confuse the simple unstructured population picture, on top of the large variation from the mean in any particular sample as already discussed. Analysis of the 20% sample population correctly identified all common ancestors.

The effects of the migrations are clearly apparent in the neutral genetic data, with the isolation of population *C* particular clear. For the mitochondrial genes at locus 1, populations *A* and *B* have the same allele for the entire population, but 27.5% of population *C* has a novel allele. This separation is even more apparent for locus 2, one of the slowly mutating loci, where 10% of population *B* differs from population *A* and all of population *C* differs from the other two populations. Populations *B* and *C* share an allele that is not present in population *A* for locus 1 on the Y chromosome: 11% and 100% of each population respectively. There is no variation at all for locus 2, the other slowly mutating locus. The two more slowly mutating loci have minimum values for all of the genetic diversity measures presented in Section 12.1.1, K_n, P_n and S_n.

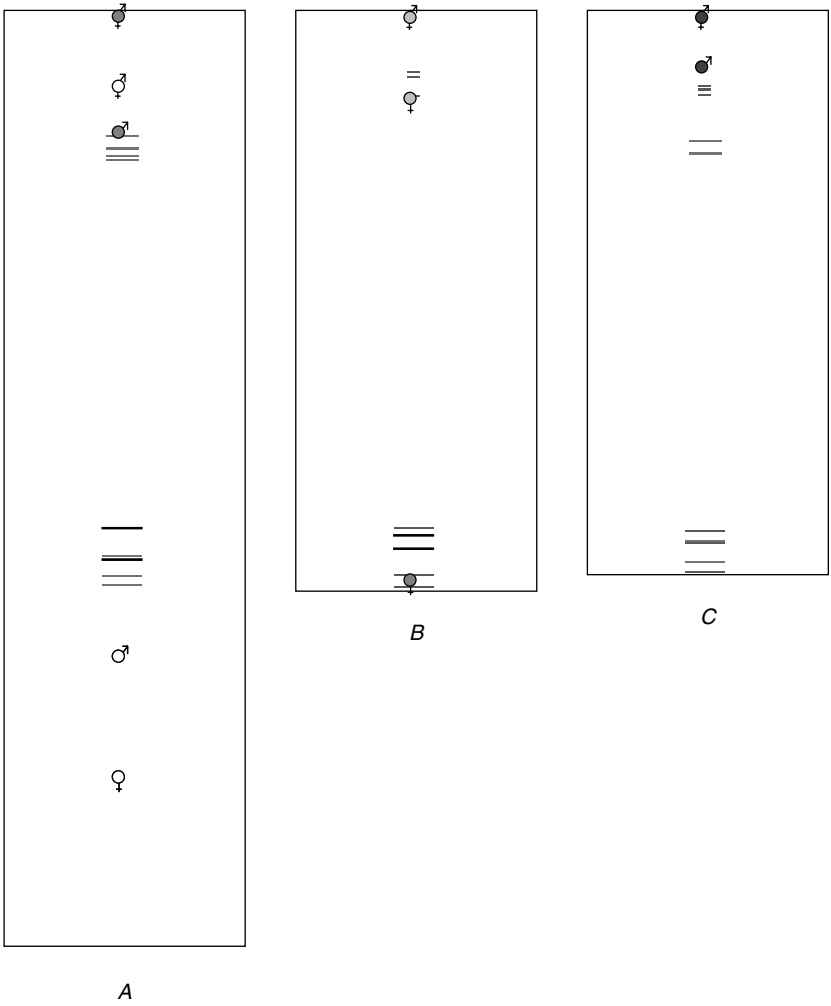


Figure 12.2. Outlines of the three simulated populations, with common ancestors and migrations indicated in the usual fashion.

The oldest population, *A*, has the most distinct alleles on the autosomes but is entirely fixed for the mitochondrial loci, and very nearly so on the Y-chromosome loci: a single individual introduced a new mutation into the population in the final generation! At autosome locus 1, every allele from populations *B* and *C* is also present in population *A*, plus two additional ones. At autosome locus 2, populations *A* and *B* have the same allele as their most

frequent allele, with greater than 90% of the population carrying it in each case. This particular allele is also in population *C*, but is only carried by less than 10% of the population. The most frequent allele in population *C*, again carried by more than 90% of the population, is novel to population *C*. This allele arose in population *C* after the replacement migrations, and subsequently had no way to spread to the other populations. The picture is similar for locus 1, although this allele arose during the replacement migrations and did in fact spread to population *B*, but did not persist there. Figure 12.3 illustrates this, showing the distribution of the two mtDNA locus-2 alleles and the two autosome locus-2 alleles discussed.

The most advantageous genotype, AA, rapidly became fixed whenever introduced to a population. In population *A* it grew quickly to 100% of the population, and although it was knocked back a little by the migrations from population *B*, it quickly re-established itself. Because the A allele was introduced into population *A* after the replacement migrations between generations 300 and 350, it did not reach either of the other populations until the male-dominated migrations from population *A* after generation 725. However, by generation 765 more than half of both populations *B* and *C* had genotype AA; by generation 794, 100% of members across all three populations were of this genotype.

Averaging 50 simulations with these settings led to the common ancestry results shown in Table 12.2. The time to the most recent common ancestor, and its associated variance, is clearly influenced by the male-dominated migrations from population *A* and the female-dominated migrations from population *B*. Both the population-*A* male genealogy and the population-*B* female genealogy have a most recent common ancestor in agreement with the time indicated by standard coalescent theory, given their respective effective population sizes of 33 males and 100 females. The other four single-population male and female most recent common ancestor times all occur approximately 100 generations further back, with a much larger variance, because of the migration-introduced lineages.

The overall biological common ancestor is ten times further back than in the single-population cases, but note that it only needs the male-dominated migrations to be established. In contrast, the overall male and female common ancestors both rely on the replacement migrations to be established, but occur at approximately the same time once these happen. The variance of the time to the overall single-sex most recent common ancestors is significantly less than the migration-affected individual population cases. These results illustrate how the effects of population structure dominate the genealogy, and thus indicate how the genealogy can carry important evidence of the migration history of the overall population.

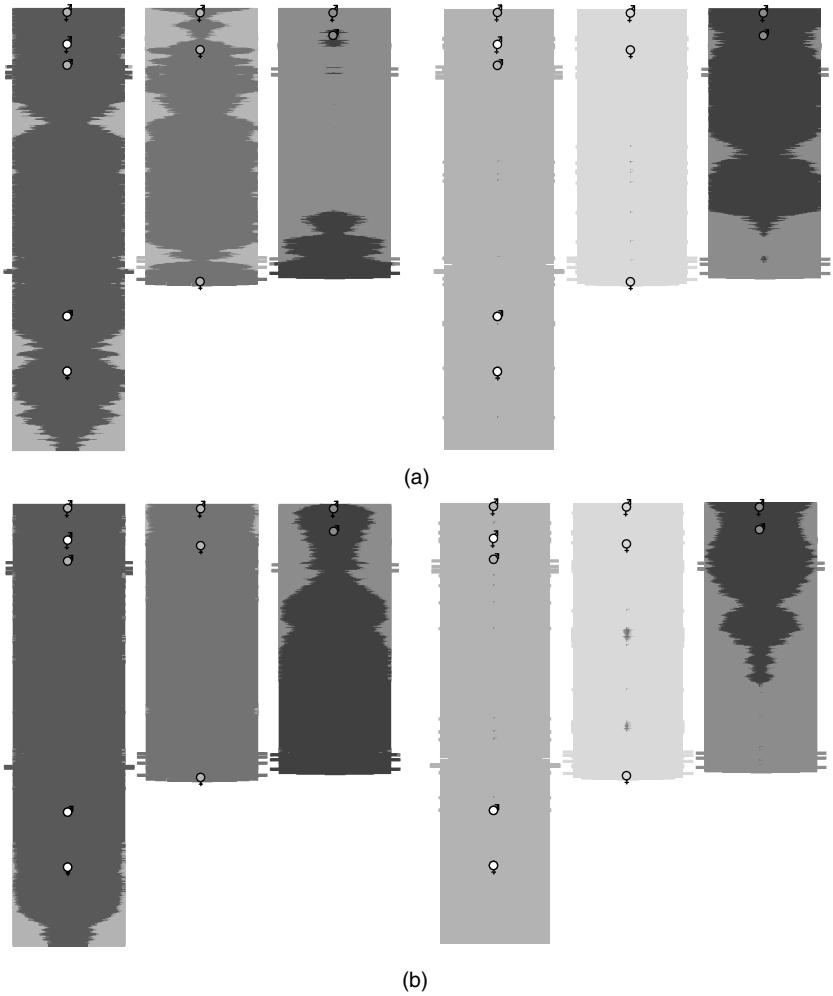


Figure 12.3. The distribution of two particular alleles within each population is shown for mitochondrial locus 2 (a) and autosome locus 2 (b). In each case the figure on the left shows the distribution of one of these alleles across the three populations, and that on the right shows the distribution of the other. The effect of the relative isolation of population *C* on the genetics is apparent by comparing the left- and right-hand figure in each case. For example, from the left portion of (a) it may be seen that the mitochondrial locus-2 allele that has historically dominated populations *A* and *B* was only prevalent in population *C* for a brief initial period, and when it reappeared as a result of later migrations, it quickly died out.
In contrast, the right-hand portion of the figure shows how the allele that dominates population *C* has only ever occurred in that same population. A similar situation is apparent in (b) for an allele at autosome locus 2.

Table 12.2. *Average common-ancestor times and locations for the polygyny simulation with migration and selection*

Population	Generation back (s_{N-1})	Location				Migrations found
		A	*B*	*C*	none	
Paternal genealogy						
All	496.4 (115.4)	98%	2%	0%	0%	3.3 (3.3)
A only	72.2 (29.7)	100%	0%	0%	0%	0.8 (0.0)
B only	176.5 (211.4)	28%	72%	0%	0%	1.9 (0.9)
C only	172.3 (194.6)	26%	2%	72%	0%	1.8 (0.7)
Maternal genealogy						
All	495.8 (142.3)	78%	16%	4%	2%	5.4 (4.9)
A only	291.6 (206.1)	82%	16%	2%	0%	3.1 (1.6)
B only	182.6 (121.2)	6%	94%	0%	0%	2.3 (0.6)
C only	275.8 (210.6)	30%	22%	48%	0%	3.4 (1.8)
Biological genealogy						
All	64.72 (13.19)	92%	8%	0%	0%	26.2 (3.5)
A only	6.74 (0.44)	100%	0%	0%	0%	15.7 (0.0)
B only	6.68 (0.47)	0%	100%	0%	0%	18.9 (0.0)
C only	6.62 (0.49)	0%	0%	100%	0%	22.6 (0.0)

The common-ancestor locations in Table 12.2 also illustrate the role of the migrations and the different effects they have on each population. For example, because very few males migrate to population *A* on the necessary time scale, the paternal population-*A* common ancestor is from population *A* 100% of the time.

The population-*B*- and population-*C*-only results for the paternal genealogy both have roughly a 3 : 1 ratio of own population to population *A* location. This is because, with an effective population of 33 breeding males, the expectation from coalescent theory for the time to the most recent common ancestor is 66 generations back, i.e. right at the time of the male migrations. This picture is not as clear for the maternal genealogy because the greater effective population means that the time scale is greater and easily encompasses both the male- and female-dominated migrations; because of the large variance, it will often reach back to the time of the replacement migrations. One result of this is the fact that the population-*C* female most recent common ancestor is in fact more often away from population *C* than on it.

The overall paternal common ancestor is almost always from population *A*, specifically in 98% of the runs, and it might seem that this is mostly because of the migration of males from population *A*. However, the fact that the time of both overall single-sex ancestors is tied to the time of the replacement migrations, as discussed above, gives an indication that this is not the full

Table 12.3. *Genetic diversity measures from the 50 simulation average of the polygyny simulation with migration and selection*

K_n is the number of distinct alleles, P_n is the average pairwise difference, and S_n the number of segregating sites (see Section 12.1.1). The full population was a constant 200 individuals in each population, with a 50 : 50 sex ratio, and the sample size was 20% of this population.

		Autosome		mtDNA		Y chromosome	
Population	Statistic	locus 1	locus 2	locus 1	locus 2	locus 1	locus 2
Sample population							
	K_n	3.78	3.38	1.04	1.24	1.24	1.02
A only	P_n	1.58	0.79	0.02	0.06	0.07	0.00
	S_n	4.28	2.84	0.04	0.24	0.24	0.02
	K_n	3.96	3.02	1.08	1.22	1.32	1.08
B only	P_n	1.62	0.65	0.03	0.09	0.11	0.03
	S_n	4.42	2.56	0.08	0.28	0.38	0.10
	K_n	4.06	3.96	1.08	1.32	1.24	1.12
C only	P_n	1.66	0.93	0.02	0.14	0.11	0.04
	S_n	4.46	3.58	0.08	0.40	0.30	0.12
	K_n	8.14	7.12	1.36	2.36	2.50	1.32
All	P_n	1.82	1.11	0.11	0.55	0.62	0.10
	S_n	7.16	6.08	0.36	1.62	1.78	0.32
Full population							
	K_n	4.54	4.12	1.16	1.44	1.38	1.08
A only	P_n	1.57	0.77	0.02	0.05	0.06	0.01
	S_n	4.90	3.50	0.16	0.44	0.38	0.08
	K_n	4.80	3.80	1.20	1.50	1.38	1.10
B only	P_n	1.62	0.66	0.02	0.08	0.10	0.02
	S_n	5.04	3.20	0.20	0.56	0.44	0.12
	K_n	4.84	4.66	1.20	1.46	1.42	1.16
C only	P_n	1.64	0.92	0.02	0.13	0.10	0.04
	S_n	5.00	4.14	0.20	0.54	0.48	0.16
	K_n	10.26	9.12	1.72	2.94	2.84	1.44
All	P_n	1.81	1.10	0.11	0.53	0.61	0.09
	S_n	8.42	7.44	0.72	2.16	2.12	0.44

picture. Furthermore, the most recent female common ancestor for the overall population is also from population *A* in 78% of the runs, and in only 16% of runs on population *B*. This shows that the nature of population *A* as the source population is the most important aspect, and that the time of the earlier replacement migrations is more significant than the time of the single-sex-dominated migrations, especially in the female case given the greater effective size of the female population.

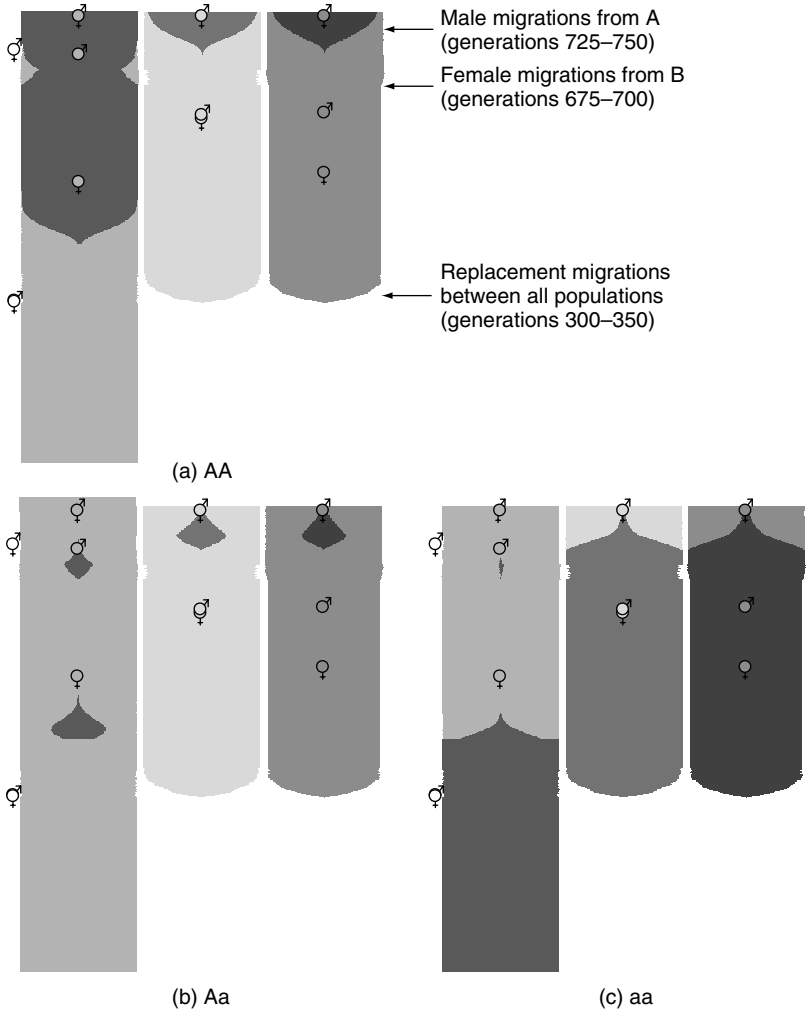


Figure 12.4. Fifty-run average genotype distributions for the polygyny simulation with migration and selection. Allele A is favoured ($s = 0.2$) and semi-dominant ($h = 0.5$).

The analysis using the sample population captures the same pattern as the full analysis, with respect to both common ancestor time and location.

The average genetic diversity statistics are shown in Table 12.3. Using the values for the sample population, all three of the mutation-rate estimators

presented in Section 12.1.1 capture the different mutation rates at the two Y-chromosome loci and the two mitochondrial loci quite well. The simulations ran with a factor of 5 difference in the mutation rate in each case. This ratio was estimated by using Tajima's formula (Equation 12.7) to be 5.0 from the mitochondrial sample and 6.2 from the Y-chromosome sample; by using Watterson's formula (Equation 12.3) to be 4.4 from the mitochondrial sample and 5.4 from the Y-chromosome sample; and by using Ewens' formula (Equation 12.5) to be 4.0 from the mitochondrial sample and 5.0 from the Y-chromosome sample.

The average genotype distributions are shown in Figure 12.4, and are as expected given the selective advantage of allele A ($s = 0.2$) and its dominance ($h = 0.5$). Genotype AA quickly dominates in population A, and then in the others when introduced into the other populations via the male-dominated migrations between generations 725 and 750. Allele a is reintroduced into population A by the female-dominated migrations from population B, leading to a reappearance of genotypes aa and Aa. However, the dominance of allele A means that allele a quickly disappears from population A (persisting a little longer in Aa heterozygotes because A is only semi-dominant) and has nearly vanished from the other two populations by the end of the simulation.

13 *Discussion*

The impact of the basic parameters of population size limits, breeding patterns, sex ratio, chance of reproduction and expected number of offspring, as well as some population-wide external influences on a single, unstructured population, was studied in the simulations presented in Chapter 10. In Chapter 11, these results were extended to include migration, and in Chapter 12, selective advantage. In all cases, the simulation results were analysed with respect to both one-parent and two-parent ancestry.

These results are summarised in the following sections, after which there is a brief discussion of their implications for understanding the origin and evolution of modern humans. Finally, the direction of some ongoing and proposed future simulation development is discussed.

13.1 Single-population summary

13.1.1 Constant demographics

The initial simulations studied involved a purely monogamous population, where all individuals were able to mate and fertility was such that a constant population was easily maintained (for a given population limit of 200 individuals). Increases in fertility beyond this level had no effect on the genealogy, and the results matched the theoretical expectations from coalescent theory for the time back to the most recent paternal and maternal common ancestors (Kingman, 1982a,b; Hudson, 1990).

The most recent biological common ancestor, i.e. the common ancestor where lineages are traced back simultaneously through both parents, was much more recent than in the single-sex cases, occurring on average only $\log_2 N$ generations back for a constant effective population of size N, with very low variance. Furthermore, a steady-state percentage of prior individuals ancestral to members of the current population was reached very quickly in the biological ancestry case.

When the analysis was carried out using only a sample of m individuals, the true most recent common ancestor for the single-sex genealogies was found approximately as expected from the $(m-1)/(m+1)$ rule (Saunders *et al.*, 1984). In contrast, the sample analysis for the most recent

biological common ancestor was much less successful at finding the true ancestor, but because the biological lineages mix so rapidly it was never wrong by more than a few generations. When the sample incorrectly identified the most recent common ancestor in the single-sex case, it was usually wrong by many generations, because the last few lineages mix very slowly – recall that the coalescence of the final two lineages takes, on average, half the overall coalescence time.

The next set of simulations involved three different fertility rate and mating chance settings, specifically TFR and reproduction chance pairs of 4/100%, 8/50% and 12/33% for the females in the population. Each of these resulted in the same total number of offspring, but with a very different effective population, and this was reflected in the time to the most recent common ancestor for each of the three kinds of genealogy studied. The results still matched coalescent theory, with an effective population equal to the census population, half the census population, and one third of the census population, respectively. Similarly, the percentage of earlier individuals ancestral to members of the current generation was reduced in proportion to the reproduction chance values.

Different mating patterns also result in different effective populations, and again the results closely matched the theoretical expectations. For example, when the mating pattern was polygyny, the chance of any two individuals sharing a father was greater than that of their sharing a mother, and therefore the paternal most recent common ancestor was, on average, more recent. The adult sex ratio was of crucial importance for these runs: when it did not match the sex ratio within the mating groups, a percentage of individuals of the sex that was over-represented in the population did not participate in the mating groups, further altering the effective population. However, even when the whole population could mate, there still remained a significant difference in the survivor percentages between the different mating patterns, reflecting the different effective populations.

A more subtle effect was introduced with the disallowing of consanguineous matings, specifically matings between siblings or cousins, except when the only alternative was population extinction (e.g. when the total population dropped below 15 individuals). This reduced the effective population by reducing the choice of mates. The impact of this change increased when the reproduction chance was lower, and was also more significant for the maternal ancestry in polygynous mating than in monogamy. For example, the case of a constant population of size 100, a TFR of 12 and a 33% chance of reproduction per female was unstable because of random variations in offspring numbers and sex ratio, often resulting in extinction. However, as the population became larger the consanguineous mating restriction became relatively less important, and the constant population was more readily maintained. The clearest

difference between cases where consanguineous mating was and was not allowed was actually in the biological genealogy, with a significant difference observed in all cases compared.

Another more subtle variation to the basic case was due to introducing the possibility of an individual that had already mated participating in a second mating group (or even more). With this possibility available to 25% of males and 10% of females, a significant change from the pure monogamy results was seen in all three common ancestry and survivor analyses: paternal, maternal and biological.

13.1.2 Varying demographics

After having covered the above mentioned basic cases, and shown that the simulation agreed with the theoretical expectations from coalescent theory, the focus moved to more complicated single-population simulations where the population size and the other demographic settings varied over time. The first set of scenarios studied were those in which the population size changed with time. Specifically, simulations were run for various exponentially growing populations (with the growth possibly restricted to specific periods), with bottlenecks at different times, and random fluctuations over the course of the simulation. These situations had very different population profiles, but often produced very similar outcomes for the common-ancestor analysis. The lineage profiles, in general, maintained more information about the differences between the cases, but these differences were often small and would be difficult to confirm in a real population.

Bottlenecks were found to have an even stronger impact on the survivor percentages than on the population size. For example, in the simulations of a constant 200 individual, purely monogamous population, with a bottleneck between 250 and 300 generations back from the final generation, not only did the population drop by 60% during the bottleneck, but the lineage-survival percentage dropped from around 70% to 40%, thus magnifying the effect in terms of ancestry. This situation was typical of all the bottleneck cases.

The random fluctuations also had a drastic effect on both the rate of lineage-merging and the survival percentage. Again with the basic constant population size, purely monogamous case as an example, with fluctuations producing a maximum population size reduction of around 50%, and an average population size reduction of only 6.6%, the survivor percentage dropped by more than a quarter, and the most recent common-ancestor times were on average around 60 generations more recent in the single-sex cases, and more than half a generation more recent for the biological ancestor. Although this second number may

sound small, it is highly significant; in fact, the two fluctuation cases were the only ones to show any significant change in the time to the most recent biological common ancestor out of all the varying population simulations.

Some of these runs were repeated for polygynous and polyandrous populations, in both cases with mating groups of four individuals, and similar effects were seen. Because of the difference in effective population between males and females, and thus common-ancestor times for the basic cases with these mating patterns, some more interesting interactions were possible. For example, with the polygynous population, the middle and late bottlenecks had a different effect on the two single-sex genealogies because of their quite different time scales. They each acted to bring the average generation for the two common ancestors closer together, because the bottleneck had more impact on the earlier ancestor and pushed it forward more strongly. For the constant population polygyny run with a late bottleneck, the most recent maternal common ancestor was brought forward nearly 200 generations on average, compared with a change of just under 50 generations in the paternal common ancestor for the same case.

Variation in mating and offspring settings over the course of a simulation was also studied, by looking at the effect of changing between two different monogamous states, one with TFR = 4 and reproduction chance 100%, and one with TFR = 8 and reproduction chance 50%, and two different 1 male/3 female polygyny states, one with an adult sex ratio of 25% male and the other with an adult sex ratio of 50% male. Simulations were carried out where the change between any two chosen states was over either 800, 400, or only 200 generations. The significant factors in determining the influence of these changes were the degree of difference between the start and the end states, indicating just how much room there was for variation, and also the depth of the genealogy for the end state, since depth gives an indication of the time scale for the coalescence of lineages, and thus how likely the end state was to be altered by the earlier state.

In no case did the biological common-ancestor generation show a significant difference from that of the end state for any of the varying cases studied, because over the $\log_2 N$ generations time scale of their mixing to a common ancestor the biological lineages of the current population did not see any significant demographic change.

13.2 Migration summary

For the simulations of migrations between three populations, only a few different cases were considered because of the difficulty of completely covering

the vast array of possibilities available to the simulation. Furthermore, less emphasis was placed on average results for these runs, because the smoothing inherent in the averaging process tended to obscure many features of particular interest.

These results showed an even greater diversity of most recent common-ancestor timings than the single-population simulations, because of the addition of extra randomness from the migrations. The variance in the average timings for each of the individual populations was very large, whereas for the overall values it was less so, but these were often very strongly affected by the migrations, especially when they were sex-specific and restricted in time. The effect was most noticeable in the biological common-ancestor analysis, where the mixing in the individual populations proceeded very rapidly as was seen in the single-population simulations, but the final, overall mixing of lineages was often delayed, having to wait until the lineages subsequently began to move between the populations.

The single-sex genealogies contained only a small number of the total migrations, because they consisted of only a small number of lineages for most of their existence. The biological genealogies, on the other hand, contained a far more complete representation of the historical migrations, because of the high steady-state level of lineage persistence.

Not only were the common ancestors in these genealogies occurring at different times, but migrations resulted in their frequently being in different populations as well – on average up to 70% of the time for the overall common ancestors. Obviously single-sex-dominated migrations from different populations were quite important in this regard, but even when the migrations involved both sexes equally, random effects ensured that the ancestor locations were spread across the different populations.

The success of the sample analysis in finding true common ancestors was higher for the migration runs, most notably for the most recent biological common ancestor. In fact, for the monogamy migration simulations discussed in Section 11.2.1, the biological-ancestor sample success rates were only 3%, 9% and 8% for each individual population, but overall the sample population analysis correctly located the true most recent common ancestor in every one of the 100 simulations. In the migration simulations presented in Section 11.1.2, the success rate of the sample population for the overall biological ancestor was 89%, and the true maternal common ancestor was found in 64 of the 65 runs where this common ancestor existed.

A simulation in which a bottleneck occurred in one population, just before an important period of migration, showed how demographic effects were highly significant in the migration runs (see Section 11.2.2). The reduction in lineages caused by the bottleneck resulted in a more closely related set

of migrants over the following generations, thus leading to a more recent common ancestor than would otherwise have been the case.

13.3 Genetics summary

The final simulations presented were an attempt to illustrate how the many features of the genealogy simulation could be brought together to study quite complicated interactions between the underlying models. Specifically, three populations with mating-pattern polygyny were simulated, with periods of migration between them. The earliest migrations were replacement migrations across all three populations, then quite near the end of the simulation were two periods of sex-specific migrations; the first female-dominated from population *B*, and the second male-dominated from population *A*.

Neutral genetics was included by allowing mutation from an initial pool of five alleles at each locus, with one Y-chromosome locus and one mitochondrial DNA locus mutating five times more slowly than all the other loci. In addition, the genotype model was used to introduce an advantageous allele to population *A* after the conclusion of the replacement migrations and well before the sex-specific migrations began.

Among the interesting features seen in a single run with these settings was an example of lineage isolation due to population structure extending the time to the most recent common paternal ancestor for one of the populations well back in time, to a time much earlier than the corresponding female common ancestor despite the mating pattern being polygyny. Also seen were several examples of novel alleles being generated and persisting on population *C*, owing to its relative isolation.

The 50 run average showed how the overall single-sex lineages, on average, relied on the initial replacement migrations to reach their final coalescence. Related to this is the high degree to which the founder population dominated the average coalescent trees. In addition, the differences in mutation rate were well reflected in the genetic diversity statistics, based on both a sample and the full population.

13.4 Implications for modern human origins

The simulations in Part I of this book focus on the evolution of species, and in particular, the problem of reconstructing the phylogenetic relationships between current and fossil species, and of current species with each other.

In contrast, the simulations in Part II are more concerned with the random processes that have resulted in an observed population, and the interplay and relative importance of the various underlying demographic and genetic parameters.

Estimating these parameters is crucially important for tracing back lineages, and the results of the genealogy simulations in this book clearly show the problems in assuming that current demographic conditions apply also to earlier times, since regular and random variations in population size, as well as variations in mating pattern, fertility and related factors have a major influence on the underlying genealogies. Furthermore, because in a single-sex genealogy the rate of lineage coalescence is maximum during the first few generations back in time, these genealogies alone contain little information about the earlier demographics of a population.

In contrast, biological (i.e. two-parent) lineages both mix extremely rapidly, and quickly reach a steady-state number of lineages, rather than showing the continued reduction to a single lineage as in the single-parent case. Thus biological lineages contain far more information about the historical demographics of a population, but are in themselves much more complicated and have no simple genetic analogue. Instead, a suite of autosomal genes must be studied to try to capture features of the biological genealogy. The rapid mixing and high percentage of surviving lineages also imply that genealogies from different autosomal genes will differ quickly and drastically. In short, different trees can easily lead to different conclusions.

The extremely high variance predicted by standard coalescent theory remains the case for more complicated simulations, and thus the validity of applying average results to a true population, which is essentially a single run, is actually very difficult to justify. This is complicated even further by migrations, especially migrations dominated by one particular sex and those restricted in time. These will have a substantial impact on common ancestry because of the constraints they place on the movement of lineages between populations, and even though migrations of a given type may have been occurring over long periods, and involving a large percentage of the population, the fact that most migrations are not evident in the genealogy of the surviving lineages will act to mask much of the evidence of these periods.

However, on the positive side, the migrations make the sample analysis generally more accurate, because much more is happening inside the individual populations before their mixing, and thus the lineages that finally do mix between the populations will generally be represented in the sample population history.

In general, the implications of simulations such as those possible through software such as *Genie* are extremely wide-ranging. For example, the ability to

simulate migrations that vary in nature, degree and timing can cast light on the impact of complicated lineage admixture, and this can be further combined with changes of breeding pattern and sex ratio, and the different impacts on mtDNA, Y-chromosome and autosome ancestry studied. In addition, issues relating to symmetry, or the lack thereof, between migrations and back-migrations, similar to the species study in Section 6.3.2, can provide many fruitful areas of simulation. Similarly, the impact of variation in other demographic parameters, such as the timing of periods of population growth with respect to the timing of migrations, has important implications for the study of human populations.

A general aim is to identify particular genetic and/or demographic signatures in the current population that can reveal aspects of the various migrations and other demographic changes responsible for the generation of that population. This has particular relevance to the recent African origin vs. multiregional evolution controversy, briefly discussed in Section 1.2. For the historical human population, there have undoubtedly been periods of population fragmentation and bottlenecks, e.g. as the result of changes towards cold and dry climates. Conversely, warmer periods and other geographic events can lead to periods of population dispersal. By constructing populations and migrations that correspond to Africa, Europe and Asia, for example (along the lines of the hominoid evolution simulations in Section 6.2), specific lineage-mixing (Nordborg, 2001b) and evolutionary geographic hypotheses may be simulated and tested (Lahr and Foley, 1998).

The genetics models in the simulation combine to provide direct means to simulate and study many interesting problems relating to the understanding and interpretation of genetic data. Specifically, there is the question of the accuracy and robustness of the various methods for estimating mutation rates in the presence of selection, migration, and population fluctuation, as well as the effect on genetic diversity of different mutation rates across loci. This last point in particular is highly significant, owing to its implications for the estimation from genetic data of the time to the most recent common ancestor.

Detecting recombination and estimating the recombination rate are also of great importance, and simulation provides a means to address these issues in a variety of different contexts. Related to this is of course the study of linkage disequilibrium and allelic frequencies, and the simulation similarly provides a platform for the study of patterns of linkage disequilibrium under situations of various mutation and recombination rates, demographic regimes and migrations.

Because population analysis is backward-looking, and many different paths lead to the same or similar patterns of genetics and genealogy, repeated numerical simulation is an excellent tool for looking for indications in the final

population that will enable differentiation between these paths. For example, in many situations there is substantial difficulty in separating the effects of selection and migration (Nordborg, 2001a). The genealogy and genotype models combine to allow many aspects of this problem to be studied,

Also benefiting from numerical simulations are general studies on the accuracy of coalescent theory for more realistic and complicated underlying population models, and also when the sample size approaches the population size. The study of such questions is well handled by simulations such as those presented in this book.

Finally, by concentrating on the demographics and genetics, it is easy to understate the importance of being able to visualise the many underlying and interacting processes. Such visualisation is central to the simulations presented in this book, and simply being able to see effects and interactions can be an invaluable aid to both teaching and research.

13.5 Future work

The genealogy simulation presented in Part II provides a highly configurable and adaptable framework for the study of ancestry and migrations across many different scenarios. In its current form, essentially all the basic demographic ingredients are provided, and so future enhancements will consist of relatively minor refinements to the nature and configuration of the various settings, possibly major enhancements to the algorithms employed to enable more efficient operation and thus the simulation of larger populations over longer times, and, most importantly, the addition of more sophisticated genetic models.

Enhancements of the first kind include allowing some variation in the constitution of the mating groups for the particular polygamous cases, perhaps by specifying an average group size rather than the current situation of a fixed mating group size specification. In addition, more precise specification and modelling of the population reduction disasters, possibly adding a sex dependence, would be useful, as would enhancing the migration specification to include different settings for the various source and destination population combinations, as the species simulation from Part I currently allows. Clearly, a more systematic study of migration scenarios than provided in Chapter 11 is required, and these migration enhancements in particular would enable finer control over this further work. Many important variations can be studied, such as moving migration windows and source–sink migrations (as for the species simulation in Chapter 6), along with many subtleties of sex-dependent migrations, and the role of selective advantage.

Turning now to the genetics models, the addition of extra autosome loci and allowing intragenic recombination will allow more refined study of linkage disequilibrium and recombination generally. This is a straightforward expansion of the existing model, but places more demands on computer resources. Also, it would be interesting to include some of the algorithms for estimating the time to the most recent common ancestor from genetic data, such as the methods of Tavaré *et al*. (1997) and Tang *et al*. (2002). The results could then be immediately compared both with each other and with the true values from the genealogy. Streamlining the export of the simulated genetic data is also appropriate for enabling more complicated and general analysis by using other tools.

Perhaps the most ambitious aim for the simulations presented in this book is their eventual combination into a simulation that can provide a unified picture of genealogical and phylogenetic processes, with implications for reconstruction on each of population, subspecies and species time scales. Although currently the simulations are quite different, with enhancements to the models, plus improved algorithms and the inevitable improvements in available computing power, their combination should provide a far more powerful and interesting simulation than either independently.

References

Abramowitz, M. and Stegun, I., eds. (1970). *Handbook of Mathematical Functions, With Formulae, Graphs, and Mathematical Tables*. New York: Dover.

Adcock, G., Dennis, E., Easteal, S. *et al*. (2001). Mitochondrial DNA sequences in ancient Australians: Implications for modern human origins. *Proc. Natl. Acad. Sci. USA* **98**(2), 537–42.

Arnason, U., Gullberg, A., Burgette, A. S. and Janke, A. (2000). Molecular estimates of primate divergences and new hypotheses for primate dispersal and the origin of modern humans. *Hereditas* **133**, 217–28.

Asfaw, B., White, T., Lovejoy, O., Latimer, B., Simpson, S. and Suwa, G. (1999). *Australopithecus garhi*: a new species of early hominid from Ethiopia. *Science* **284**, 629–35.

Avise, J. (2000). *Phylogeography*. Cambridge, MA: Harvard University Press.

Ayala, F. (1995). The myth of Eve: Molecular biology and human origins. *Science* **270**, 1930–6.

Ayala, F. and Escalante, A. (1995). The evolution of human populations: A molecular perspective. *Molec. Phylogenet. Evol*. **5**(1), 188–201.

Ayres, K. and Balding, D. (2001). Measuring gametic disequilibrium from multilocus data. *Genetics* **157**, 413–23.

Barbujani, G. and Bertorelle, G. (2001). Genetics and the population history of Europe. *Proc. Natl. Acad. Sci. USA* **98**, 22–5.

Beerli, P. and Felsenstein, J. (1999). Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. *Genetics* **152**, 763–73.

Begun, D. (2002). European hominoids. In W. Hartwig, ed., *The Primate Fossil Record*. Cambridge: Cambridge University Press, pp. 339–68.

(2003). Planet of the apes. *Scient. Am.*, August 2003, pp. 74–83.

Behrensmeyer, A., Todd, N., Potts, R. and McBrinn, G. (1997). Late Pliocene faunal turnover in the Turkana basin, Kenya and Ethiopia. *Science* **278**, 1589–94.

Brookfield, J. (1997). Importance of ancestral DNA ages. *Nature* **388**, 134.

Brunet, M., Guy, F., Pilbeam, D., *et al.* (2002). A new hominid from the Upper Miocene of Chad, Central Africa. *Nature* **418**, 145–51.

Cann, R. (2001). Genetic clues to dispersal in human populations: Retracing the past from the present. *Science* **291**, 1742–8.

Cann, R., Stoneking, M. and Wilson, A. (1987). Mitochondrial DNA and human evolution. *Nature* **325**, 31–5.

Chang, J. (1999). Recent common ancestors of all present-day individuals, with Discussion and Reply. *Adv. Appl. Prob.* **31**, 1002–38.

Collard, M. and Wood, B. (2000). How reliable are human phylogenetic hypotheses? *Proc. Natl. Acad. Sci. USA* **97**(9), 5003–6.

Corless, R., Gonnet, G., Hare, D., Jeffrey, D. and Knuth, D. (1996). On the Lambert W function. *Adv. Comp. Math.* **5**, 329–59.

Cracraft, J. (1983). Species concepts and speciation analysis. In R. H. Johnston, ed., *Current Ornithology*, vol. 1. New York: Plenum Press, pp. 159–87.

Cronquist, A. (1987). A botanical critique of cladism. *Bot. Rev.* **53**, 1–52.

Cummins, J. (1999). Evolutionary forces behind human infertility. *Nature* **397**, 557–8.

Darwin, C. (1859). *The Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. New York: Doubleday.

(1871). *The Descent of Man and Selection in Relation to Sex*. New York: Modern Library Reprint.

Day, W., Johnson, D. and Sankoff, D. (1986). The computational complexity of inferring rooted phylogenies by parsimony. *Math. Biosci.* **81**, 33–42.

Donnelly, P. and Tavaré, S. (1995). Coalescents and genealogical structure under neutrality. *A. Rev. Genet.* **29**, 401–21.

Dorit, R., Akashi, H. and Gilbert, W. (1995). Absence of polymorphism at the ZFY locus on the human Y chromosome. *Science* **268**, 1183–5.

Ewens, W. (1972). The sampling theory of selectively neutral alleles. *Theor. Popul. Biol.* **3**, 87–112.

Falush, D., Stephens, M. and Pritchard, J. (2003). Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics* **164**, 1567–2587.

Farris, J. (1970). Methods for computing Wagner trees. *Syst. Zool.* **19**, 83–92.

Felsenstein, J. (1982). Numerical methods for inferring evolutionary trees. *Q. Rev. Biol.* **57**(4), 379–404.

(1993). PHYLIP (phylogeny inference package). Distributed by the author. Department of Genetics, University of Washington, Seattle. http://evolution.genetics.washington.edu/phylip.html

Fu, Y. (1994). A phylogenetic estimator of effective population size or mutation rate. *Genetics* **136**, 685–92.

(1998). Probability of a segregating pattern in a sample of DNA sequences. *Theor. Popul. Biol.* **54**, 1–10.

Fu, Y. and Li, W. (1993). Statistical tests of neutrality of mutations. *Genetics* **133**, 693–709.

(1996). Estimating the age of the common ancestor of men from the ZFY intron. *Science* **272**, 1356–7.

(1997). Estimating the age of the common ancestor of a sample of DNA sequences. *Molec. Biol. Evol.* **14**, 195– 9.

(1999). Coalescing into the 21st century: An overview and prospects of coalescent theory. *Theor. Popul. Biol.* **56**, 1–10.

Gage, T. (1998). The comparative demography of primates: With some comments on the evolution of life histories. *A. Rev. Anthropol.* **27**, 197–221.

Gagneux, P., Wills, C., Gerloff, U. *et al*. (1999). Mitochondrial sequences show diverse evolutionary histories of African hominoids. *Proc. Natl. Acad. Sci. USA* **96**, 5077–82.

Gebo, D., MacLatchy, L., Kityo, R., Deino, A., Kingston, J. and Pilbeam, D. (1997). A hominoid genus from the early Miocene of Uganda. *Science* **276**, 401–4.

Gee, H. (2001). Return to the planet of the apes. *Nature* **412**, 131–2.

Gibbons, A. (1998). Calibrating the mitochondrial clock. *Science* **279**(5347), 28–9.

Goldstein, D., Linares, A. R., Cavalli-Sforza, L. and Feldman, M. (1995). Genetic absolute dating based on microsatellites and the origin of modern humans. *Proc. Natl. Acad. Sci. USA* **92**, 6723–7.

Goodman, M. (1995). Epilogue: A personal account of the origins of a new paradigm. *Molec. Phylogenet. Evol.* **5**(1), 269–85.

Goodman, M., Porter, C., Czelusniak, J., *et al*. (1998). Toward a phylogenetic classification of primates based on DNA evidence complemented by fossil evidence. *Molec. Phylogenet. Evol.* **9**(3), 585–98.

Graham, R. and Foulds, L. (1982). Unlikelihood that minimal phylogenies for a realistic biological study can be constructed in reasonable computational time. *Math. Biosci.* **60**, 133–42.

Graham, R., Knuth, D. and Patashnik, O. (1994). *Concrete Mathematics: A Foundation for Computer Science*, 2nd edn. Boston: Addison-Wesley.

Griffiths, R. (1999). The time to the ancestor along sequences with recombination. *Theor. Popul. Biol.* **55**, 137–44.

Griffiths, R. and Tavaré, S. (1994). Ancestral inference in population genetics. *Stat. Sci.* **9**, 307–19.

(1996). Monte Carlo inference methods in population genetics. *Math. Comput. Modelling* **23**, 141–58.

Groves, C. (2001). *Primate Taxonomy*. Washington, DC: Smithsonian Institution Press.

Haile-Selassie, Y. (2001). Late Miocene hominids from the Middle Awash, Ethiopia. *Nature* **412**, 178–81.

Hammer, M. (1995). A recent common ancestry for human Y chromosomes. *Nature* **378**, 376–8.

Harding, R. (1996). New phylogenies: An introductory look at the coalescent. In P. Harvey, A. Brown, J. M. Smith and S. Nee, eds., *New Uses for New Phylogenies*. New York: Oxford University Press, pp. 15–22.

Harding, R., Fullerton, S., Griffiths, R., *et al*. (1997). Archaic African and Asian lineages in the genetic ancestry of modern humans. *Am. J. Hum. Genet.* **60**, 772–89.

Harpending, H., Batzer, M., Gurvan, M., Jorde, L., Rogers, A. and Sherry, S. (1998). Genetic traces of ancient demography. *Proc. Natl. Acad. Sci. USA* **95**, 1961–7.

Harpending, H., Sherry, S., Rogers, A. and Stoneking, M. (1993). The genetic structure of ancient human populations. *Curr. Anthropol.* **34**(4), 483–96.

Hartl, D. and Clark, A. (1997). *Principles of Population Genetics*, 3rd edn. Sunderland, MA: Sinauer Associates.

Hayes, B., Visscher, P., McPartlan, H. and Goddard, M. (2003). Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome Res.* **13**, 635–43.

Hedrick, P. (1987). Gametic disequilibrium measures: Proceed with caution. *Genetics* **117**, 331–41.

Hennig, W. (1966). *Phylogenetic Systematics*. Urbana, IL: University of Illinois Press.

Hey, J. and Wakeley, J. (1997). A coalescent estimator of the population recombination rate. *Genetics* **145**, 833–46.

Hrdy, S. (2000). The optimal number of fathers: Evolution, demography and history in the shaping of female mate preferences. *Annals NY Acad. Sci.* **907**, 75–96.

Hudson, R. (1983). Properties of a neutral allele model with intragenic recombination. *Theor. Popul. Biol.* **23**, 183–201.

(1990). Gene genealogies and the coalescent process. In D. Futuyma and J. Antonovics, eds., *Oxford Surveys in Evolutionary Biology*. Oxford: Oxford University Press, pp. 1–44.

(1998). Island models and the coalescent process. *Molec. Ecol.* **7**, 413–18.

Hull, D. (1997). The ideal species concept – and why we can't get it. In M. Claridge, H. Dawah and M. Wilson, eds., *Species: The Units of Biodiversity*. London: Chapman and Hall, pp. 357–80.

Huxley, T. (1863). *Evidence as to Man's Place in Nature*. New York: Appleton.

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921.

Jobling, M. and Tyler-Smith, C. (1995). Fathers and sons: The Y chromosome and human evolution. *Trends Genet.* **11**(11), 449–56.

Jones, S., Martin, R. and Pilbeam, D., eds. (1991). *The Cambridge Encyclopedia of Human Evolution*. Cambridge: Cambridge University Press.

Jorde, L. (1995). Linkage disequilibrium as a gene-mapping tool. *Am. J. Hum. Genet.* **56**, 11–14.

Jorde, L., Bamshad, M. and Rogers, A. (1998). Using mitochondrial and nuclear DNA markers to reconstruct human evolution. *BioEssays* **20**, 126–36.

Kaplan, H. (1994). Evolutionary and wealth flows theories of fertility: Empirical tests and new models. *Popul. Devel. Rev.* **20**(4), 753–91.

Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature* **217**, 624–6.

Kingman, J. (1982a). Exchangeability and the evolution of large populations, *in* G. Koch and F. Spizzichino, eds., *Exchangeability and Probability in Statistics*. New York: North-Holland, pp. 97–112.

(1982b). On the genealogy of large populations. In J. Gani and E. Hannan, eds., *Essays in Statistical Science*. Sheffield, UK: Applied Probability Trust, pp. 27–43.

Köhler, M. and Moyà-Solà, S. (1997). Ape-like or hominid-like? The positional behaviour of *Oreopitheus bambolii* reconsidered. *Proc. Natl. Acad. Sci. USA* **94**, 11747–50.

Kuhner, M., Yamato, J. and Felsenstein, J. (1998). Maximum likelihood estimation of population growth rates based on the coalescent. *Genetics* **149**, 429–34.

Lahr, M. and Foley, R. (1998). Towards a theory of modern human origins: Geography, demography, and diversity in recent human evolution. *Yrbk Phys. Anthropol.* **41**, 137–76.

Lancaster, J. (1997). The evolutionary history of human parental investment in relation to population growth and social stratification. In P. Gowaty, ed., *Feminism and Evolutionary Biology*. New York: Chapman and Hall, pp. 466–88.

Leakey, M., Fiebel, C., McDougall, I. and Walker, A. (1995). New four-million-year-old hominid species from Kanapoi and Allia Bay, Kenya. *Nature* **376**, 565–71.

Leakey, M., Spoor, F., Brown, F., *et al.* (2001). New hominin genus from eastern Africa shows diverse middle Pliocene lineages. *Nature* **410**(6827), 433–40.

Lewin, R. (1993). *The Origin of Modern Humans*. No. 47 in Scientific American Library Series. New York: W. H. Freeman & Co.

Lewontin, R. (1988). On measures of gametic disequilibrium. *Genetics* **120**, 849–52.

Mann, A. and Weiss, M. (1996). Hominoid phylogeny and taxonomy: A consideration of the molecular and fossil evidence in an historical perspective. *Molec. Phylogenet. Evol.* **5**(1), 169–81.

Mayr, E. (1969). *Principles of Systematic Biology*. New York: McGraw-Hill.

Mishler, B. and Donoghue, M. (1982). Species concepts: A case for pluralism. *Syst. Zool.* **31**, 491–503.

Moore, W. (1995). Inferring phylogenies from mtDNA variation: Mitochondrial-gene trees versus nuclear-gene trees. *Evolution* **49**(4), 718–26.

Nee, S., May, R. and Harvey, P. (1994). The reconstructed evolutionary process. *Phil. Trans. R. Soc. Lond.* B **344**, 305–11.

Nei, M. (1991). Relative efficiencies of different tree-making methods for molecular data. In M. Miyamoto and J. Cracraft, eds., *Phylogenetic Analysis of DNA Sequences*. New York: Oxford University Press, pp. 90–128.

Neilsen, R. (2001). Statistical tests of selective neutrality in the age of genomics. *Heredity* **86**, 641–7.

Neuhauser, C. and Krone, S. (1997). The genealogy of samples in models with selection. *Genetics* **145**, 519–34.

Nielsen, D., Ehm, M. and Weir, B. (1998). Detecting marker-disease association by testing for Hardy-Weinberg disequilibrium at a marker locus. *Am. J. Hum. Genet.* **63**, 1531–40.

Nordborg, M. (2001a). Coalescent theory. In D. Balding, M. Bishop and C. Canning, eds., *Handbook of Statistical Genetics*. Chichester: Wiley, pp. 179–208.

(2001b). On detecting ancient admixture. In P. Donnelly and R. Foley, eds., *Genes, Fossils, and Behaviour: An Integrated Approach to Human Evolution*. NATO Science Series. Amsterdam: IOS Press, pp. 123–36.

Nordborg, M. and Tavaré, S. (2002). Linkage disequilibrium: What history has to tell us. *Trends Genet.* **18**, 83–90.

O'Connell, J. (1999). Genetics, archaeology, and Holocene hunter-gatherers. *Proc. Natl. Acad. Sci. USA* **96**, 10562–3.

Oxnard, C. (1997). The time and place of human origins: Implications from modelling. In G. Clark and C. Willermet, eds., *Conceptual Issues in Modern Human Origins Research*. New York: Aldine de Gruyter, pp. 369–91.

(2000). Morphometrics of the primate skeleton and the functional and developmental underpinnings of species diversity. In P. O'Higgins and M. Cohn, eds., *Development, Growth and Evolution: Implications for the study of the hominoid skeleton*. Linnean Society Symposium Series no. 20. London: Academic Press, pp. 235–63.

(2004). Design, level, interface and complexity: Morphometric interpretation revisited. In F. Anapol, R. German and N. Jablonski, eds., *Shaping Human Evolution*. Cambridge: Cambridge University Press, pp. 391–414.

Oxnard, C. and Wessen, K. (2001). Modelling divergence, interbreeding and migration: Species evolution in a changing world. In I. Metcalfe, J. Smith, M. Morwood and I. Davidson, eds., *Faunal and Floral Migrations and Evolution in SE Asia-Australasia*. Lisse: A. A. Balkema, pp. 373–85.

Palumbi, S., Cipriano, F. and Hare, M. (2001). Predicting nuclear gene coalescence from mitochondrial data: The three-times rule. *Evolution* **55**(5), 859–68.

Pilbeam, D. (1996). Genetic and morphological records of the Hominoidea and hominid origins: a synthesis. *Molec. Phylogenet. Evol.* **5**, 155–68.

Posada, D., Crandall, K. and Holmes, E. (2002). Recombination in evolutionary genomics. *A. Rev. Genet.* **36**, 75–97.

Pritchard, J., Seielstad, M., Perez-Lezuan, A. and Feldman, M. (1999). Population growth of human Y chromosomes: A study of Y chromosome microsatellites. *Molec. Biol. Evol.* **16**(12), 1791–8.

Raup, D., Gould, S., Schopf, T. and Simberloff, D. (1973). Stochastic models of phylogeny and the evolution of diversity. *J. Geol.* **81**, 525–42.

Relethford, J. (1998). Genetics of modern human origins and diversity. *A. Rev. Anthropol*. **27**, 1–23.

Rogers, A. and Jorde, L. (1995). Genetic evidence on modern human origins. *Hum. Biol.* **67**, 1–36.

Rook, L., Bondioli, L., Köhler, M., Moyà-Solà, S. and Macchiavelli, R. (1999). *Oreopithecus* was a bipedal ape after all: Evidence from the iliac cancellous architecture. *Proc. Natl. Acad. Sci. USA* **96**, 8795–9.

Ruvolo, M. (1996). A new approach to studying modern human origins: Hypothesis testing with coalescence time distributions. *Molec. Phylogenet. Evol.* **5**(1), 202–19.

(1997). Molecular phylogeny of the hominoids: Inferences from multiple independent DNA sequence data sets. *Molec. Biol. Evol.* **14**(3), 248–65.

Sarich, V. and Wilson, A. (1967). Immunological time scale for hominoid evolution. *Science* **158**, 1200–3.

Satta, Y., Klein, J. and Takahata, N. (2000). DNA archives and our nearest relative: The trichotomy problem revisited. *Molec. Phylogenet. Evol.* **14**, 259–75.

Saunders, I. W., Tavaré, S. and Watterson, G. A. (1984). On the genealogy of nested subsamples from a haploid population. *Adv. Appl. Prob.* **16**, 471–91.

Seielstad, M., Minch, E. and Cavalli-Sforza, L. (1998). Genetic evidence for a higher female migration rate in humans. *Nature Genet.* **20**, 278–80.

Senut, B., Pickford, M., Gommery, D., Mein, P., Cheboi, C. and Coppens, Y. (2001). First hominid from the Miocene (Lukeino formation, Kenya). *C. R. Acad. Sci. Paris* **332**, 137–44.

Sepkoski, J. Jr and Kendrick, D. (1993). Numerical experiments with model monophyletic and paraphyletic taxa. *Paleobiology* **19**(2), 168–84.

Sherry, S., Batzer, M. and Harpending, H. (1998). Modeling the genetic architecture of modern populations. *A. Rev. Anthropol.* **27**, 153–69.

Shoshani, J., Groves, C., Simons, E. and Gunnell, G. (1996). Primate phylogeny: Morphological and molecular results. *Molec. Phylogenet. Evol.* **5**(1), 102–54.

Simpson, G. (1945). The principles of classification and a classification of mammals. *Bulletin 85, American Museum of Natural History*.

(1961). *Principles of Animal Taxonomy*. Vol. 20 of *Columbia Biological Series*. New York: Columbia University Press.

Smith, S. and Harrold, F. (1997). A paradigm's worth of difference? Understanding the impasse over modern human origins. *Yrbk Phys. Anthropol.* **40**, 113–38.

Sokal, R. (1985). The continuing search for order. *Am. Nat.* **126**(6), 729–49.

Stewart, C. and Disotell, T. (1997). Primate evolution – in and out of Africa. *Molec. Biol. Evol.* **14**, 195–9.

Strauss, E. (1999). Can mitochondrial clocks keep time? *Science* **283**, 1435–8.

Swofford, D. (n.d.). PAUP*: Phylogenetic analysis using parsimony. http://paup.csit.fsu.edu/index.html

Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**, 437–60.

(1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–95.

Takahata, N. and Satta, Y. (1997). Evolution of the primate lineage leading to modern humans: Phylogenetic and demographic inferences from DNA sequences. *Proc. Natl. Acad. Sci. USA* **94**, 4811–15.

Tang, H., Siegmund, D., Shen, P., Oefner, P. and Feldman, M. (2002). Frequentist estimation of coalescence times from nucleotide sequence data using a tree-based partition. *Genetics* **161**, 447–59.

Tavaré, S., Balding, D., Griffiths, R. and Donnelly, P. (1997). Inferring coalesence times from DNA sequence data. *Genetics* **145**, 505–18.

Tavaré, S., Marshall, C., Will, O., Soligo, C. and Martin, R. (2002). Using the fossil record to estimate the age of the last common ancestor of extant primates. *Nature* **416**, 726–9.

Underhill, P., Jin, L., Lin, A. *et al.* (1997). Detection of numerous Y chromosome biallelic polymorphisms by denaturing high-performance liquid chromatography. *Genome Res.* **7**, 996–1005.

Vigilante, L., Stoneking, M., Harpending, H., Hawkes, K. and Wilson, A. (1991). African populations and the evolution of human mitochondrial DNA. *Science* **253**, 1503–7.

Wakeley, J. (1997). Using the variance of pairwise differences to estimate the recombination rate. *Genet. Res.* **69**, 45–8.

Ward, S., Brown, B., Hill, A., Kelley, J. and Downs, W. (1999). *Equatorius*: A new hominoid genus from the middle Miocene of Kenya. *Science* **285**, 1382–6.

Watterson, G. (1975). On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7**, 256–76.

Westendorp, R. and Kirkwood, T. (1998). Human longevity at the cost of reproductive success. *Nature* **396**, 743–6.

White, T., Suwa, G. and Asfaw, B. (1994). *Australopithecus ramidus*, a new species of hominid from Aramis, Ethiopia. *Nature* **371**, 306–12.

Whitfield, L., Sulston, J. and Goodfellow, P. (1995). Sequence variation of human Y chromosome. *Nature* **378**, 379–80.

Wijsman, E. and Cavalli-Sforza, L. (1984). Migration and genetic population structure with special reference to humans. *A. Rev. Ecol. Syst.* **15**, 279–301.

Wiley, E., Siegel-Causey, D., Brooks, D. and Funk, V. (1991). *The Compleat Cladist. A Primer of Phylogenetic Procedures*. Special Publication No. 19. Lawrence, KS: The University of Kansas, Museum of Natural History.

Wills, C. (1995). When did Eve live? An evolutionary detective story. *Evolution* **49**, 593–607.

Wollenberg, K. and Avise, J. (1998). Sampling properties of genealogical pathways underlying population pedigrees. *Evolution* **52**(4), 957–66.

Wood, B. (2002). Hominid revelations from Chad. *Nature* **418**, 133–5.

Wood, B. and Collard, M. (1999). The human genus. *Science* **284**, 65–71.

Wood, B. and Richmond, B. (2000). Human evolution: Taxonomy and paleobiology. *J. Anat.* **196**, 19–60.

Zapata, C., Carollo, C. and Rodriguez, S. (2001). Sampling variance and distribution of the D' measure of overall gametic disequilibrium between multiallelic loci. *Annals Hum. Genet.* **65**, 395–406.

Zondervan, K. and Cardon, L. (2004). The complex interplay among factors that influence allelic association. *Nature Revs Genet.* **5**, 89–100.

Index